Санкт-Петербургский государственный университет

Кафедра компьютерного моделирования и многопроцессорных систем

Шведов Олег Юрьевич

Выпускная квалификационная работа бакалавра

Анализ эффективности работы общественного транспорта с применением технологий больших данных

Направление 010300

Фундаментальная информатика и информационные технологии

Научный руководитель,
Ph.D.,
доцент
Корхов В. В.

Санкт-Петербург
2016

Saint Petersburg State University

Department of Computer Modelling and Multiprocessor Systems

Shvedov Oleg

Bachelor's Thesis

Efficiency analysis of public transport operations using big data technologies

Field of study 010300
Fundamental Informatics and Information Technology

Scientific supervisor,
Ph.D.,
Associate Professor
Korkhov V.V.

Saint Petersburg
2016

Содержание

Введение **5**

Постановка задачи **8**

Обзор литературы **9**

Глава I. Эффективность общественного транспорта **11**

1.1. Текущая ситуация в сфере общественного транспорта 12

1.2. Критерии эффективности общественного транспорта 12

1.3. Анализ критерия эффективности доступа 15

1.3.1 Покрытие и дистанция доступа 15

1.4. Анализ критерия эффективности доступности 16

1.4.1 Частота рейсов и количество маршрутов 18

1.5. Анализ критерия избыточности 18

1.5.1 Неэффективность расположения остановок 19

1.5.2 Локационная задача покрытия множества 20

1.5.3 Лагранжева эвристика 22

1.5.4 Задача максимального покрытия территории 24

Глава II. Методология больших данных **26**

2.1. Определение больших данных 26

2.2. Apache Hadoop 27

2.2.1 Компоненты Apache Hadoop 28

2.2.2 Распределенная файловая система HDFS 30

2.2.3 Парадигма MapReduce 31

2.2.4 Программный интерфейс Hadoop Streaming API . . . 33

2.3. Облачная платформа Microsoft Azure HDInsight 33

2.4. Большие данные и исследования общественного транспорта 35

Глава III. Реализация задачи **38**

3.1. Описание исходных данных 38

3.1.1 Транспортные данные 38

3.1.2 Административное деление и население 40
3.1.3 Улично-дорожная сеть 41
3.2. Программное обеспечение и конфигурация 43
3.3. Подготовка данных 44
3.3.1 Конвертация файла УДС 44
3.3.2 Выгрузка данных в HDFS облачного кластера 44
3.3.3 Фильтрация данных 45
3.3.4 Выполнение задания MapReduce 46
3.3.5 Перемещение данных в геоинформационную систему 48
3.4. Реализация анализа параметра доступа 49
3.5. Реализация анализа параметра доступности 52
3.6. Совмещенный анализ параметров доступа и доступности . . 53
3.7. Реализация анализа избыточности 55

Выводы 58

Заключение 59

Приложение 60
Пример содержимого GTFS-канала 60
Скрипт загрузки данных улично-дорожной сети 63
Сценарий загрузки файлов из облачного BLOB-хранилища 64

Список литературы 65

Введение

Система общественного транспорта, как одна из наиболее важных составляющих инфраструктуры городов, несомненно оказывает существенное влияние на их жизнедеятельность. Обладание досконально полной информацией о функционировании всех компонент этих систем позволяет гибко и динамично осуществлять широчайший комплекс манипулятивных действий по управлению. Это особенно полезно в условиях современных городов, в особенности, в мегаполисах с уже сложившейся структурой общественного транспорта.

Услуги, предоставляемые системами общественного транспорта, являются важным компонентом процессов общегородского планирования и управления, однако существующая идеология городского управления в области общественного транспорта остается достаточно общей и не предполагает манипулирование системой в сколь бы то ни было крупном масштабе. Затруднительным остается контроль за функционированием транспорта в режиме реального времени, сбор точных статистических данных для всей системы общественного транспорта, отсутствие унифицированного инструментария транспортного моделирования на уровне городов.

Традиционные методы сбора данных, такие как анкетирование, тематические исследования, городские аудиты, интервью, фокус-группы, этнография, имеют ряд серьезных недостатков: они генерируются на разрозненной, не непрерывной основе, имеют малое количество переменных, агрегируются до грубого масштаба и часто имеют ограниченный доступ. Коренная причина этого кроется в ограниченности выборки исходных данных, которые довольно плотно сосредоточены, дополнительно основываясь на каком-то конкретном временном и пространственном промежутке и физически ограничены по объему, а также довольно дороги для генерации и анализа. Большинство современных знаний о городах были выведены из этих несовершенных и дефицитных данных.

Одна из ветвей качественного преобразования городского управления — концепция Smart City, опирающаяся на сферу больших данных, которая в свою очередь предлагает современные подходы к сбору потока данных обес-

печивающих более сложное, масштабное и мелкозернистое информационное покрытие областей городской жизнедеятельности. Именно автоматически генерируемые данные должны предоставлять всю необходимую информацию о самых малых составляющих городской инфраструктуры, позволяя управлять ими с максимальной степенью автоматизма и в режиме реального времени средствами управляющей информационной системы. Взаимодействие таких данных и информационных систем требует минимального человеческого участия, снижая тем самым сложность управления городом, а также предоставляя значительное количество всевозможной информации, которую невозможно собрать имеющимися средствами. Таким образом, реализация концепции Smart City может упростить управление городской инфраструктурой, позволив при этом узнать гораздо больше информации о ее функционировании и состоянии.

В соответствии с концепцией, сфера городского общественного транспорта также подлежит фундаментальной реорганизации для получения возможности более качественно эксплуатировать ее. Осуществление преобразований в этой сфере невозможно без введения кардинально иных методов, таких как централизованное управление в режиме реального времени: целенаправленный сбор данных о перемещении подвижного состава и пассажиров посредством телекоммуникационного оборудования в единый центр транспортного управления позволяет отслеживать функционирование всей системы в целом, что невозможно было бы осуществить имеющимися методами, такими, как локальный ручной подсчет и визуальное наблюдение.

С исследовательской точки зрения, собираемые транспортные данные могут быть использованы для более детального изучения функционирования общественного транспорта, к примеру, для более точного моделирования, для выяснения различных статистических закономерностей, либо для стратегического планирования. Одной из характеристик функционирования, информацию о которой можно извлечь из транспортных данных, выступает эффективность общественного транспорта. В свою очередь, данная характеристика весьма полезна для целей краткосрочного и долгосрочного планирования, а также для проведения иных исследований, например, для разработки инновационных методов проектирования.

Самостоятельное исследование характеристики эффективности тем не менее может быть весьма важным для обеспечения удобства пользования общественным транспортом, что является одной из основных функций транспортных служб. Информация об эффективности общественного транспорта помогает достоверно определять районы, где требуется внести локальные или крупномасштабные изменения транспортной системы, в том числе в связке с автомобильным транспортом. С другой стороны, эффективность достаточно релевантно отражает доступность и удобство общественного транспорта для перемещений на нем жителей города.

Из этого следует, что эффективный, инновационный и качественный (комфортный) общественный транспорт привлекает большее количество пассажиров, которые выбирают его вместо использования личного автотранспорта, что позитивно сказывается на воздействии на окружающую среду, дорожном трафике, и на энергетической эффективности. Также доказана (отметим) взаимосвязь развития транспортной, в том числе общественной, составляющей региона и решения проблем его устойчивого развития [38].

Постановка задачи

Цель данной работы состоит в выполнении действий по исследованию различных характеристик эффективности функционирования общественного транспорта. Реализация данного исследования требует привлечения информационных средств и средств анализа исходных транспортных данных, которые в свою очередь должны содержать статическую информацию об инфраструктуре системы общественного транспорта и ее функционировании за какой-либо временной промежуток.

Особенностью работы является применение в ходе исследования методов и инструментов анализа больших данных для манипуляции исходными транспортными данными.

Сформулируем задачи, выполнение которых позволит достичь поставленной цели:

- Изучение предметной области:
  - Текущее состояние общественного транспорта.
  - Существующие на данный момент направления исследований в области эффективности общественного транспорта.
  - Различные формулировки критериев эффективности общественного транспорта.
- Поиск исходных данных о функционировании транспортной системы, удовлетворяющих поставленным требованиям.
- Подготовка набора транспортных данных средствами набора утилит Apache Hadoop.
- Использование полученных результативных данных для расчета требуемых параметров эффективности в геоинформационной системе.

Обзор литературы

Работа [39] раскрывает характеристики качества услуг, предоставляемых системой общественного транспорта с точки зрения восприятия их пассажирами. Источник [43] содержит обзор комплекса научных и прикладных направлений: способы и особенности обеспечения мобильности населения, организация маршрутных перевозок и оценка качества транспортного обслуживания населения. Представлен перечень показателей качества транспортного обслуживания населения общественным транспортом, определена и структурирована классификация направлений совершенствования работы городского общественного транспорта.

В работе [33] произведена попытка выделения отличительных характеристик эффективных транспортных систем с целью их перенесения в города с менее эффективным общественным транспортом. Для этого была разработана система определения эффективных транспортных систем, базирующаяся на DEA-подходе (Data Envelopment Analysis). Данный подход также фигурирует в исследовании [22].

Работы [17; 28; 29] содержат исследования, направленные на оптимальное размещение объектов остановок и рассматривают, соответственно, проблему нахождения минимального эффективного количества остановок автобусных экспресс-маршрутов, организуемых на основе имеющегося набора маршрутов и их остановок, проблему охвата маршрутной сетью урбанизированных территорий и проблему реализации комбинированного подхода к избавлению от избыточных остановок вместе с введением минимального количества новых в нуждающихся районах города.

Авторы статьи [9] описывают вариант реализации методики определения степени эффективности транспортных систем на основе искусственных нейронных сетей с многослойными перцептронов.

В статье [35] показан вариант использования транспортных данных в формате GTFS для аналитических целей: комбинация информации об автобусных остановках и передвижениях пассажиров позволила авторам выяснить истинную величину пассажиропотоков в системе экспресс-автобусов

и в системе простых автобусов, чьи участки маршрутов иногда проходят в одних и тех же местах.

Работа [2] показывает возможности применения вычислительных методик больших данных для городского управления и проведения исследовательских мероприятий по изучению функционирования различных составляющих городской жизни. Автор отмечает особую роль систем датчиков в организации и налаживании процесса сбора данных с разных источников, в том числе с подвижного состава общественного транспорта. По словам автора, этот новый тип данных способен дать толчок к изменению процесса городского стратегического планирования, превратив его из долгосрочного в гибкий краткосрочный, который охватывает любой масштаб, как локальный, так и общегородской.

Глава I. Эффективность общественного транспорта

Развитый и эффективный общественный транспорт напрямую способствует устойчивому экономическому развитию обслуживаемых им территорий и является необходимым условием обеспечения мобильности населения и доступности его точек притяжения. Кроме этого, транспортная система оказывает большое влияние на общий ритм жизнедеятельности, состояние окружающей среды, устойчивость социальной составляющей и качество жизни на покрываемых ею территориях. Безусловно, повышение степени транспортной эффективности благотворно сказывается и на данных параметрах. Неэффективная же транспортная система и связанные с ней формы городской инфраструктуры будут ограничивать экономические и социальные возможности регионов.

Основополагающую роль в развитии транспортной системы играет общественный транспорт, так как именно он способен оказать наиболее сильный положительный энергетический и экологический эффекты, такие как уменьшение уровня потребления горючего топлива, шума, заторов и загрязнения воздуха по сравнению с более высокой долей личного автотранспорта в перевозках. Развитие общественного транспорта заключается в расширении охвата транспортных услуг для наибольшего процента горожан, населяющих конкретные районы.

Необходимо разработать комплекс средств, позволяющих понять, оценить и контролировать функционирование имеющихся транспортных систем таким образом, чтобы гарантировать достижение краткосрочных и долгосрочных планировочных целей. Измерения степени эффективности работы общественного транспорта могут иметь решающее значение при оценке политических целей, а также при планировании будущих улучшений.

1.1. Текущая ситуация в сфере общественного транспорта

Рост мегаполисов в последние десятилетия повлек множество социально-экономических проблем. Многие из таких проблем напрямую касаются сферы общественного транспорта как одного из важнейших элементов эволюции и роста урбанизированных областей. Так, население, вытесняемое на окраины с дешевым жельем из центральных районов города, либо переезжающее в них из других мест проживания, обеспечивает стабильно повышенный спрос на услуги общественного транспорта, что увеличивает размеры бюджетных расходов на эксплуатацию городской инфраструктуры. По причине того, что зачастую общественный транспорт на окраинах развит более слабо по сравнению с городским центром, данный спрос оказывается еще более острым. Отчасти это вытекает из недостаточной развитости улично-дорожной сети окраинных районов и больших расстояний, которые приходится покрывать горожанам в процессе внутригородской маятниковой миграции.

В целом, неорганизованный рост городов дает сообразно неорганизованное развитие транспортной системы, в результате чего текущее функционирование транспортных систем многих городов нельзя назвать эффективным.

Безусловно, любое комплексное мероприятие по улучшению общественного транспорта невозможно без заключения рабочих соглашений между всеми вовлекаемыми органами власти, в частности, разделения обязанностей, административной координации, финансирования. Комплексный управленческий подход является необходимым фундаментом для разработки и внедрения методик повышения эффективности общественного транспорта.

1.2. Критерии эффективности общественного транспорта

Понятие эффективности функционирования общественного транспорта может быть проанализировано на основе нескольких специфичных кри-

териев, связанных с производительностью транспортных учреждений и качетсвом реализуемых ими услуг. К таким критериям относят [33]:

- Доступность какой-либо из составляющих транспортной системы, определяемая расстоянием между начальными пунктами в маршрутах пассажиров и ближайшей остановкой общественного транспорта, и расстоянием между последней остановки общественного транспорта в маршрутах пассажиров и их конечными точками назначения. Уменьшение данного расстояния повышает доступность транспортных услуг и, как следствие, увеличивает географический охват городских территорий, делая их более присвособленными для населения с точки зрения простоты перемещений.

- Время в пути, определяемое быстротой движения подвижного состава и геометрией маршрутов. Быстрота определяется как функция, зависящая от расстояния, условий дорожного движения и качества движения.

- Надежность, определяемая степенью неопределенности временного планирования. Она может быть измерена по количеству поездок и их времени по отношению к другим, совершенным с задержками и временами этих задержек.

- Частота, определяемая временными интервалами между каждым рейсом. Пассажиры должны знать расписание движения и их изменения в течение дня, в выходные или праздничные дни или во время других особых случаев.

- Максимальная нагрузка, определяемая количеством пассажиров в часы пик в зависимости от вместимости подвижного состава.

- Характеристики единиц подвижного состава, в том числе: возраст, экологичность, шум, технологические параметры (габариты дверных проемов, высота пола, доступность для пассажиров с ограниченными возможностями), показатели комфортности.

– Адекватность информации и вспомогательного оборудования: укрытия от непогоды на остановках, информационные табло с картами и расписанием движения, хорошо различимые указатели остановок и транспортных средств.

– Мобильность в соответствии с потребностями, то есть, планирование набор маршрутов общественного транспорта таким образом, чтобы обеспечивать наибольшую степень покрытия необходимой площади и наибольшую гибкость при выборе подходящего маршрута пассажирами.

Помимо указанных требований, к эффективности работы общественного транспорта относятся также показатели производительности, такие как: стоимость эксплуатации, оптимальное количество единиц подвижного состава и обслуживающего персонала, которое не влияет на текущий уровен качества предоставляемых транспортных услуг, отношение количества пассажиров, пользующихся общественным транспортом к общей численности населения, проживающего на заданной территории, длина маршрутной сети по отношению к общей длине улично-дорожной сети заданной территории, уровень удовлетворенности пассажиров.

В настоящей работе будем рассматривать эффективность с двух точек зрения:

1) Простота пользования услугами общественного транспорта, возможность использования системы на основе ее близости — *доступ (access)*;

2) Сокращение общего времени поездки, пригодность транспортной сети в качестве средства перемещения от точки старта до точки назначения за разумный промежуток времени, оперативность функционирования системы — *доступность (accessibility)*.

В свою очередь, их тоже можно разбить на частные определения. Понятие доступа рассматривается как близость общественного транспорта для населения со стремлением к уменьшению расстояния, которое требуется преодолеть человеку для доступа к ближайшей остановке/станции общественного транспорта в предположении, что остальные характеристики,

такие как время поездки, удовлетворительны. С другой стороны, близость общественного транспорта прямым образом влияет в том числе и на общее время поездки от начального пункта до пункта назначения. Также в рамках понятия доступа используется понятие охвата территории сетью общественного транспорта. Понятие доступности рассматривается как возможность осуществления перевозок транспортной системой, существование возможности добраться от стартовой точки до точки назначения наиболее оптимальным способом: число пересадок, скорость движения, регулярность обслуживания или частота рейсов.

Необходимо отметить, что эти понятия тесно взаимосвязаны и зависят друг от друга в том случае, когда система успешно работает и используется пассажирами.

1.3. Анализ критерия эффективности доступа

Предполагается, что увеличение доступности общественного транспорта приводит к более активному его использованию, в том числе, вместо личных автомобилей. Как уже было сказано, данная миграция оказывает большое количество положительных эффектов на обшегородском уровне.

Данный показатель отвечает за близость или стоимость использования транспортных услуг [21], причем наиболее выгодно и эффективно на этот показатель можно влиять посредством расположения остановочных пунктов общественного транспорта максимально удовлетворительным для населения образом. Оптимизация расположения остановочных пунктов является часто обсуждаемой проблемой [14; 32], так как расположение не в последнюю очередь влияет на скорость доступа, время и общее удобство поездки. Безусловно, необходимо стремиться к наиболее оптимальному и удобному расположению остановок на заданной территории.

1.3.1. Покрытие и дистанция доступа

Исследуется доступность от места проживания; в качестве исходной информации выступает перепись городского населения по всем районам города. Доступность места назначения, а также других точек притяжения

требует значительно более глубокого изучения маршрутов передвижения пассажиров и моделирования пассажиропотоков.

Измерение доступа населения к инфраструктуре общественного транспорта может выражаться в оценке близости ближайшей остановки общественного транспорта на территории заданного района по отношению к стандартной удовлетворительной пешей дистанции или времени прогулки.

В среднем, для случая городских автобусов, максимальной удовлетворительной дистанцией, которую нужно преодолеть до ближайшего остановочного пункта для 90% населения, проживающего на конкретной городской территории или в городе в целом, считается 400 метров [12], однако она может выбираться динамическим образом с учетом конкретных обстоятельств и условий местности. Исходя из этого, определим так называемые буферные зоны доступности, представляемые окружностями с центрами в точках остановок и радиусом 400 метров.

Тогда можно вычислить минимальное расстояние из любой точки на территории рассматриваемого района до ближайшей остановки. Если эти расстояния не превышает радиус буферной зоны доступности данной остановки, то район считается имеющим удобный доступ к общественному транспорту. В данной работе выполняется сбор центроид районов в качестве представителей соответствующих областей, от которых с помощью евклидовой метрики измеряется дистанция до остановок общественного транспорта.

1.4. Анализ критерия эффективности доступности

Доступность является одним из ключевых вопросов транспортного планирования и землепользования в целом. Это понятие описывает легкость достижения необходимых/желаемых точек притяжения [19]. По существу, доступность означает способность обеспечить доставку пессажира до желаемых точек притяжения, однако в силу многогранности понятия, имеются проблемы в его изучении. Выделяют несколько способов использования понятия доступности:

– Удобство (не дистанция) доступа

– Распределение транспортного влияния

– Варианты проезда

– Последовательность транспорта

– Взаимосвязь с политическими институтами

– Воздействия новых разработок

– Планирование пасажирских и бизнес перевозок

По утверждению [4] доступность связана с качествами транспортной системы (например, скорость движения транспортных средств), а также с качеством системы землепользования (например, плотность и типы). В то же время, это понятие имеет прямое отношение к экономическому (доступ к рабочим местам работников, клиентов к местам расположения различных услуг, социальные контакты), а также к экологическому секторам.

Оценка функционирования общественного транспорта с точки зрения доступности может помочь в решении вопросов его зачастую невыгодного положения в вопросах перевозок населения и его собственного распределения. Доступность существенно зависит от проектирования объектов транспортной инфраструктуры, таких как маршруты общественного транспорта, остановки, улично-дорожная сеть, а также наличия в непосредственной близости различных точек притяжения. Также имеется зависимость от таких свойств, как правдивость расписаний движения общественного транспорта и ощущение безопасности в пути.

Крайне важно выделить необходимые пункты назначения и предоставить в непосредственной близости от этих пунктов соответствующие транспортные услуги с удовлетворительным временем в пути до них [20]. Соответственно, не менее важна разработка модели, способной измерить доступности сети общественного транспорта и точек притяжения. Кроме того, в рамках этих моделей необходимы стандарты доступности, позволяющие однозначно устанавливать планировочные критерии землепользования.

1.4.1. Частота рейсов и количество маршрутов

Выполнение условия наибольшего покрытия городской территории системой общественного транспорта не всегда означает его максимальную привлекательность для горожан. В связи с этим требуется разработка и использование дополнительных стратегических подходов при анализе эффективности транспортных услуг.

Существует ряд методик по оценке параметра доступности, таких как изохронный подход [30]. В работе [27] исследуется скорость движения подвижного состава. В данной работе этот параметр оценивается с точки зрения частоты совершаемых рейсов для остановок общественного транспорта. Напомним, что параметр частоты рейсов оказывает прямое влияние на эффективность общественного транспорта и на составное время в пути от начальной точки до конечной точки притяжения. С другой стороны, частота относится к производительности самой системы общественного транспорта, и к удобству пользования ее услугами.

Данный параметр, как и параметр доступа, предлагается проецировать на городской район в целом, а именно, рассчитывать среднее значение от всех рейсов, совершенных на всех остановках, лежащих в пределах данного района на всех маршрутах, которые каждая из остановок обслуживает. Для этого будет произведена обработка транспортных данных с целью формирования формата данных, пригодного для такого расчета.

1.5. Анализ критерия избыточности

Повышение эффективности работы общественного транспорта в условиях сложившихся транспортных систем имеет особую экономическую и политическую специфику. В частности, более выгодным и легко реализуемым направлением действий оказывается изучение степени избыточности общественного транспорта для дальнейшего избавления от нее. Так как проблема избыточности неразрывно связана с проблемами, характерными для параметра доступа и ее решения направлены на те же цели, что и в случае доступа, то в рамках данной работы будем изучать избыточность

точно так же с точки зрения неэффективности расположения остановочных пунктов.

Уменьшение количества остановок на маршрутах положительно сказывается на общей скорости передвижения транспортных средств и их эксплуатационной скорости [26], несмотря на ограничения, связанные с сохранением общего покрытия как минимум на прежнем уровне. С другой стороны, размещение лишних остановок практически не дает положительного эффекта с точки зрения как доступности, так и производительности общественного транспорта, а лишь увеличивает нагрузку на систему в целом и снижает качество обслуживания [29].

Скорость подвижного состава также можно повысить за счет обустройства частично или полностью выделенных полос на дорогах или за счет интеграции с более скоростным легкорельсовым общественным транспортом. Тем не менее, именно уменьшение количества остановок играет большую роль в первостепенной оптимизации существующей транспортной системы, не в последнюю очередь за счет более низкой стоимости проведения соответствующих действий. Другой положительный эффект заключается в уменьшении эксплуатационных издержек транспортной инфраструктуры за счет сохранения минимально требуемого количества остановочных пунктов, обслуживаемых эффективными маршрутами.

Безусловно, наилучших результатов можно достичь только совмещением процессов удаления избыточных элементов транспортной системы только и увеличения производительности и доступности ее сохраняемой части.

1.5.1. Неэффективность расположения остановок

Процесс измерения степени неэффективности расположения остановок на заданной территории целесообразно проводить после измерения степени покрытия этой территории общественным транспортом. Неэффективной считается остановка, попадающая в области доступа других остановок: для остановок i и j имеет место неэффективность расположения остановки i в том случае, если для соответствующих множеств областей доступа N_i и N_j имеет место $N_i \subset N_j$.

Случаи неэффективного расположения остановок не редки даже при соблюдении городскими службами фиксированного расстояния между ними, так как городские стандарты предписывают выполнение этого условия для соседних остановок только одного маршрута или одной улицы. К примеру, если на двух параллельных улицах с расстоянием между ними в 200 метров проходят разные автобусные маршруты, то по городскому стандарту их остановки должны быть расположены вдоль каждой из улиц не чаще, чем раз в 400 метров, однако возможна ситуация, когда буферная зона доступности остановки с одной улицы будет пересекаться с буферной зоной доступности остановки на другой улице.

Таким образом, анализ избыточности с точки зрения неэффективности расположения остановок также оказывается важной стратегической мерой эффективности, кроме упомянутых. Для этого надо оценивать существующую конфигурацию остановочных пунктов.

1.5.2. Локационная задача покрытия множества

Локационная задача покрытия множества (Location Set Covering Problem, LSCP) — формализация подхода к измерению степени избыточности существующего транспортного комплекса с точки зрения размещения остановочных пунктов в пределах заданной территории. Позволяет определить, насколько возможно повышение производительности функционирования сети общественного транспорта при сокращении количества остановочных пунктов на маршрутах [17; 37].

Постановка LSCP для измерительной системы избыточности использует следующие обозначения:

- i — индекс обслуживаемой зоны (района), принадлежащей множеству всех обслуживаемых зон I ($i \in I$);
- j — индекс текущей остановки, принадлежащей множеству всех существующих остановок J ($j \in J$);
- d_{ij} — кратчайшее расстояние или время в пути между любой точкой зоны i и остановкой j;

– S — стандартная дистанция доступа или время в пути;

– $N_i = \{j \mid d_{ij} \leqslant S\}$;

– $x_j = \begin{cases} 1, & \text{если остановка } j \text{ остается в зоне обслуживания} \\ 0 & \text{в противном случае;} \end{cases}$

Тогда задача LSCP имеет вид:

$$\text{Минимизировать} \quad Z = \sum_j x_j \tag{1}$$

при ограничениях:

$$\sum_{j \in N_i} x_j \geqslant 1 \quad \forall i \in I, \tag{2}$$

$$x_j \in \{0, 1\} \quad \forall j \in J. \tag{3}$$

То есть требуется минимизировать количество остановочных пунктов, необходимых для обеспечения полного покрытия заданной обслуживаемой области. Ограничение (2) указывает, что что каждая область должна обслуживаться по крайней мере одним остановочным пунктом. Это гарантирует, что все районы в настоящее время обслуживаются существующей конфигурацией остановок и будут продолжать обслуживаться меньшим количеством остановочных пунктов. Ограничение (3) гарантирует целочисленность переменных решения. Таким образом, необходимо принимать решения в определении сохраняемых остановочных пунктов, содержащихся в текущей сети общественного транспорта.

Данная задача представляет собой пространственный вариант задачи о покрытии множества [13; 31]. Разница между этими двумя задачами заключается в форме целевой функции. Задача о покрытии множества включает взвешивание переменных решения в целевой функции ($\sum_j a_j x_j,\ a_j \geqslant 0$). В LSCP же вес каждого участника равен единице.

Задача LSCP трудно решаема в смысле оптимальности для среднего и крупного масштаба; активно изучаются эвристические методики ее решения [6; 25]. Например, эвристика лагранжевой релаксации оказалась эффективной для крупномасштабных задач [18]. Ниже приводится описание данного метода.

1.5.3. Лагранжева эвристика

Изначально данная эвристика была направлена на решение стандартной задачи покрытия множества, поэтому рассмотрим ее формулировку.

Пусть заданы множества $I = \{1 \dots m\}$ и $J = \{1 \dots n\}$, а также семейство подмножеств $I\{P_j,\ j \in J\}$. Множество I, множество J и индикаторный вектор множества P_j могут быть рассмотрены как соответственно множество строк, множество столбцов и вектор-столбец бинарной матрицы.

Подмножество $C \subseteq J$ называется *формой покрытия*, если $\bigcup_{j \in C} P_j = I$.

Пусть c_j — стоимость столбца j. Тогда задача покрытия множества — это нахождение минимального стоимостного покрытия матрицы. Задача может быть формализована в виде задачи бинарного линейного программирования:

$$\text{Минимизировать} \sum_{j \in J} c_j x_j \tag{4}$$

при ограничениях:

$$\sum_{i \in P_j} x_j \geqslant 1, \quad \forall i \in I,$$

$$x_j \in \{0, 1\}, \quad \forall j \in J$$

и где:

$$x_j = \begin{cases} 1, & \text{если } j \text{ принадлежит покрытию,} \\ 0 & \text{в противном случае.} \end{cases}$$

Очевидно, задача (4) имеет решение тогда и только тогда, когда $\bigcup_{j \in J} P_j = I$; также она является NP-полной задачей [16].

Для комбинаторной задачи оптимизации Лагранжева эвристика состоит из трех этапов:

1) Выбор релаксации для задачи

2) Использование субградиентного методы для решения Лагранжевой двойственности

3) Установление на каждой итерации текущего решения релаксируемой задачи в качестве целесообразного решения исходной задачи

Релаксируя ограничения задачи (4), получаем Лагранжеву двойственность в ней — задачу LD:

$$\max_{\pi \in \mathbb{R}_+} W(\pi), \tag{5}$$

где для заданного π значение $W(\pi)$ является оптимальным решением следующей задачи $R(\pi)$:

$$\text{Минимизировать } \min_{j \in J} \left(c_j - \sum_{i \in P_j} \pi_i \right) x_j + \sum_{i \in I} \pi_i, \tag{6}$$

$$\text{где} \quad x_j = \{0, 1\}, \quad \forall j \in J.$$

Очевидно, что проблема (5) имеет то же решение, что и непрерывная версия (4).

В данном случае нас интересует лишь третий этап эвристики. Пусть $\{x_j^*, \quad j \in J\}$ является оптимальным решением (6), а также задано $C = \{j \in J \mid x_j^* = 1\}$. Если C не является покрытием, то удаляем множество $\bigcup_{j \in C} P_j$ покрываемых строк нашей матрицы и действуем в соответствии с жадным эвристическим принципом [8] для покрытия оставшихся строк введением новых столбцов.

Столбец k ($k \in C$) является *избыточным*, если $C \setminus \{k\}$ является покрытием. Покрытие является *избыточным*, если оно содержит избыточный столбец. Предположим, что покрытие C является избыточным. Тогда покажем, как улучшить его: пусть $R = \{k \in C \mid \bigcup_{j \in C \setminus (k)} P_j = I\}$ — множество избыточных столбцов. Тогда мы можем найти наименьшую стоимость покрытия, которую можно извлечь из C путем решения особой версии задачи о покрытии множества:

$$\text{Минимизировать } \sum_{j \in R} c_j x_j \tag{7}$$

при ограничениях:

$$\sum_{\{j \in R \mid i \in P_j\}} x_j \geqslant 1, \quad i \in \bigcup_{j \in C \setminus R} (I \setminus P_j),$$

$$x_j \in \{0, 1\}, \quad \forall j \in R.$$

Пусть T — оптимальное покрытие данной задачи. Тогда наименьшая стоимость покрытия, которая может быть извлечена из C, представляет собой $(C \setminus R) \cup T$.

1.5.4. Задача максимального покрытия территории

Одним из ограничений цели решения задачи (1) является требование сохранения обслуживаемых общественным транспортом территорий таковыми. Данное условие не всегда оказывается уместным: при проектировании может потребоваться исключать конкретные городские районы из общего числа обслуживаемых транспортом. Причиной этого бывают низкие объемы пассажиропотоков или чрезмерно высокие показатели эксплуатационных расходов.

Задача максимального покрытия территории (Maximal Covering Location Problem, MCLP) ослабляет данное требование [7]. Для определения этой задачи в дополнение к начальным переменным локационной задачи покрытия множества введем следующие обозначения: пусть a_i — текущий/ожидаемый пассажиропоток в районе i ($i \in I$), p — число выбираемых остановок,

$$y_i = \begin{cases} 1, & \text{если район } i \text{ имеет удовлетворительную доступность,} \\ 0 & \text{в противном случае.} \end{cases}$$

Тогда задача максимального покрытия территории формулируется как задача максимизации общей доли населения, в настоящее время имеющего доступ к общественному транспорту и которое будет продолжать его иметь:

$$\text{Максимизировать} \quad \sum_i a_i y_i$$

при условиях:

$$\sum_{j \in N_i} x_j \geqslant y_i \quad \forall i \in I, \tag{8}$$

$$(b): \quad \sum_j x_j = p, \tag{9}$$

$$(c): \quad x_j = y_i = \{0, 1\} \quad \forall i, j. \tag{10}$$

Условие (8) позволяет учитывать, какие обслуживаемые территории покрыты остановками, выбранными для сохранения. Условие (9) указывает, что выбрано ровно p таких остановок. Условие (10) накладывает ограничение на целочисленность переменных решения.

Таким образом, задача максимального покрытия локации направлена на выявление городских районов, являющихся наиболее привлекательными для их сохранения в качестве обслуживаемых общественным транспортом. Дополнительное весомое преимущество применения задачи максимального покрытия перед локационной задачи покрытия множества заключается в еще большей экономии бюджетных средств за счет большего числа удаляемых остановочных пунктов: удаление неэффективных городских районов из сети общественного транспорта влечет удаление их нее всех расположенных на их территории остановок, не ограничиваясь сохранением хотя бы минимального их количества.

Глава II. Методология больших данных

Большие данные, или технологии интенсивных данных — это новая технологическая тенденция в области науки, промышленности и бизнеса, тесно связанная практически со всеми аспектами человеческой деятельности. Возможности современных технологий, таких как облачные вычисления, а также повсеместное развитие сетевой инфраструктуры обеспечивают устойчивую платформу для автоматизации всевозможых процессов сбора хранения, обработки и визуализации данных. Приспособившись к работе с постоянно растущими объемами данных, современная наука может предложить промышленной сфере новаторские методы научного анализа, а промышленность способна создать передовые и быстро развивающиеся технологии и инструменты больших данных для научных и общественных целей [11].

2.1. Определение больших данных

– «Новое поколение технологий и архитектур, предназначенных для извлечения экономической выгоды из очень больших объемов самых разнообразных данных, позволяя быстро их получать, изучать и/или анализировать» [15].

– «Значительный объем структурированных или неструктурированных данных, который настолько велик, что его обработка с использованием традиционных программных технологий и баз данных затруднительна» [36].

– «Крупномасштабные, быстро накапливающиеся и разнообразные информационные активы, которые требуют экономически эффективных, инновационных форм обработки информации для более глубокого понимания предметной области с целью принятия каких-либо решений» [5].

– Зачастую понятие 5V расширяют дополнительными характеристиками, отражающими возможность преобразования данных в течение их

жизненного цикла и необходимость связывания исходных данных с обработанными, сохраняя их смысловую ценность: Динамичность/Изменчивость (Dynamicity/Variability) и Связность (Linkage).

2.2. Apache Hadoop

Apache Hadoop [24; 40; 53] — программная экосистема с открытым исходным кодом, образующая каркас, пригодный для разработки и выполнения распределенных приложений, обрабатывающих значительные объемы данных. Hadoop разработан в рамках распределенных систем как альтернатив мейнфреймам, но в отличие от других распределенных систем направлен не на выполнение значительной вычислительной работы, а на обработку большого объема данных, перемещение которых оказывается слишком неэффективным. Поэтому в основе Hadoop лежит идея перемещения кода выполнения к данным, но не наоборот, то есть клиентские машины посылают на кластер исполняемые программы, после чего Hadoop перемещает их на те машины кластера, где в результате распределения данных оказались требуемые для обработки блоки данных.

Преимущества Hadoop по сравнению со стандартными инструментами распределенных вычислений:

– Доступность — выполнение Hadoop возможно на кластерах, собираемых из стандартных машин и комплектующих, объединенных в сеть, физически расположенных в одном месте, либо в рамках инфраструктуры облачных вычислений.

– Надежность — по причине стандартности оборудования архитектура Hadoop разработана с учетом частых отказов, которые обрабатываются таким образом, чтобы при их возникновении характеристики кластера ухудшались плавно и постепенно.

– Линейная масштабируемость — при увеличении объема данных и вычислений достаточно добавить новые узлы в кластере.

– Простота — программируемая высокоуровневая среда позволяет создавать параллельный эффективный код.

По отношению к классическим реляционным системам управления базами данных (РСУБД) Hadoop имеет ряд отличительных особенностей:

– Масштабирование по горизонтали вместо вертикального: в классических РСУБД расширение объемов данных влечет установку более мощного сервера для их обработки; увеличение объема ресурсов кластера Hadoop сводится к добавлению новых машин.

– Использование пар ключ/значение вместо реляционных таблиц. Нередки ситуации, когда даные плохо структурируются реляционными таблицами или слишком велики для такой структуризации. Основная единица данных в Hadoop — пара ключ/значение.

– Функциональное программирование вместо декларативных запросов. Для составления исполняемых в Hadoop программ требуется самостоятельно описать необходимые шаги обработки данных вместо написания запросов SQL. Однако имеется инструментарий компиляции SQL-подобных запросов в исполняемый код.

– Автономная пакетная обработка вместо оперативных транзакций и использования B-деревьев, то есть Hadoop не предназначен для произвольного считывания и обновления нескольких записей; вместо этого оптимальное использование Hadoop предполагает однократную запись и многократное чтение данных, хранящихся пакетно.

2.2.1. Компоненты Apache Hadoop

Программно Hadoop представляет собой набор процессов-демонов на различных серверах кластерной сети. Перечислим категории демонов:

1) *NameNode* — главный демон-диспетчер пространства имен файловой системы HDFS, распределяющий низкоуровневые задачи ввода/вывода между демонами DataNode и ведущий учет разбиения файлов на блоки. Также он поддерживает дерево файловой системы и метаданные всех файлов/каталогов. Под этот демон выделяется отдельный не дублируемый сервер.

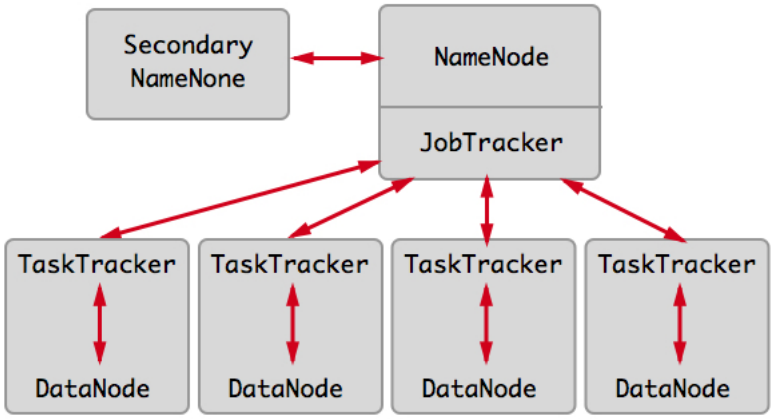

Рисунок 1 — Структурные элементы Hadoop

2) *DataNode* — демон подчиненных машин, отвечающий за считывание и запись блоков HDFS в физические файлы в локальной файловой системе, получая информацио о нахождении блоков файлов от NameNode. Взаимодействует с другими DataNode, осуществляя репликацию блоков данных. Периодически передают на NameNode информацию о состоянии и локальных изменениях.

3) *Secondary NameNode* — вспомогательный демон, осуществляющий мониторинг состояния кластера HDFS. Под этот демон также выделяется отдельный не дублируемый сервер. В отличие от NameNode не получает и не протоколирует информацию об изменениях HDFS в режиме реального времени, а только взаимодействует с ним для создания снимком метаданных HDFS.

4) *JobTracker* — демон, строящий общий план выполнения задач. Он определяет, какие файлы требуется обработать в рамках выполнения задачи, назначает задачам узлы и следит за ходом их выполнения. Выделяется на обособленный сервер, играющий роль главного узла кластера.

5) *TaskTracker* — множество демонов, подчиненных JobTracker и управляющие исполнением отдельных назначенных задач на конкретных узлах кластера. TaskTracker периодически оповещает JobTracker о ходе выполнения задач.

Общая схема взаимодействий между компонентами Hadoop представлена на рис. 1.

2.2.2. Распределенная файловая система HDFS

HDFS (Hadoop Distributed File System) — файловая система, предназначенная для хранения очень больших объемов распределенных данных и с потоковой схемой доступа к ним в кластерах обычных машин [34].

Свойства HDFS:

- Хранение очень больших файлов.
- Потоковый доступ к данным: следование концепции однократной записи и многократного чтения даст максимальную эффективность обработки данных на HDFS.
- HDFS, как и Hadoop в целом, не требует дорогостоящего оборудования с высокой надежностью, так как система спроектирована для работы на кластерах, состоящих из стандартных машин и с учетом высокой вероятности отказов отдельных узлов.
- Высокая пропускная способность в HDFS предпочтительнее скорости доступа.

Одна из основных концепций HDFS — блочность хранимых файлов данных (стандартный размер блока 64 Мбайт). Блоки одного файла можно хранить распределенно, что используется в том числе для репликации, а абстракция блока вместо абстракции файла упрощает подсистему хранения (фиксированный размер позволяет легко вычислять количество блоков, которое может вместить диск) и организацию метаданных. Безопасность данных в HDFS достигается механизмом репликации метаданных узла NameNode в нескольких файловых системах и использования Secondary NameNode для создания контрольных точек. Высокая доступность обеспечивается использованием двух узлов NameNode в конфигурации «активный/резервный». Для этого оба NameNode должны использовать общее хранилище данных с высокой доступностью для хранения журнала изменений. Узлы DataNode,

в свою очередь, должны отправлять отчеты обоим NameNode. Переходом от активного узла NameNode к резервному управляет так называемый контроллер преодоления сбоев.

2.2.3. Парадигма MapReduce

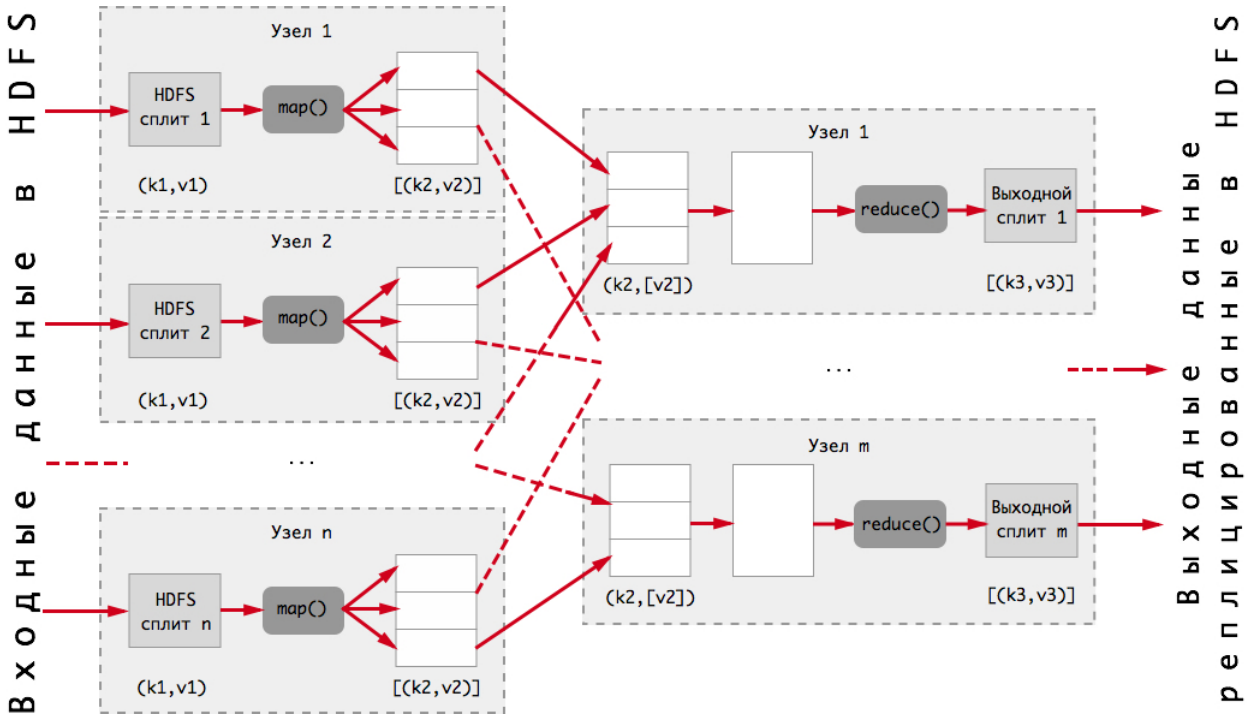

Рисунок 2 — Схема работы MapReduce

Одна из важнейших функций Hadoop — запуск программ MapReduce.

MapReduce [10] — это программная модель распределенной обработки больших объемов данных. Задание MapReduce представляет собой единицу работы, которую надо выполнить для клиента, и состоит из набора входных данных, программы MapReduce и конфигурационной информации.

Примитивы обработки данных в MapReduce — это *распределители* и *редукторы*, для которых определяются соответствующие функции обработки. Примитивы собственно данных — списки и пары ключ/значение.

Для выполнения задания MapReduce принимает входные данные этого задания, разбивает их на сплиты фиксированного размера и создает задачи распределения, содержащие входные данные в виде списка пар ключ/значение `[(k1, v1)]`.

На фазе распределения функция-распределитель (`map()`) принимает на вход отдельную задачу распределения и выполняет определенную последовательность действий над каждой парой `(k1, v1)` во входном сплите данных, преобразуя ее в список пар `[(k2, v2)]`. Задачи распределения выполняются по возможности на тех же узлах, на которых входные данные хранятся в HDFS (принцип оптимизации локальности данных). Выходные данные задач распределения записываются на локальный диск машины, а не в HDFS, так как по причине локальности этих данных реплицировать их в HDFS невыгодно.

Для масштабирования применяется функция разбивки, определяющая, каким редукторам посылать выработанные в распределителях пары ключ-значение. Выбор зависит от конфигурации расположения данных на кластерах и от хеширования ключа. При наличии нескольких функций-редукторов задачи распределения разделяют свои выходные данные на разделы. В каждом разделе может быть много ключей, но все записи для заданного ключа находятся в одном разделе — так называемое тасование. Каждая задача редукции таким образом получают свои входные данные от многих распределителей. Для минимизации объема передаваемых между распределителями и редукторами данных также возможно определение функции-комбинатора, сокращающей вывод распределителя.

На фазе редукции обрабатываются промежуточные результаты, полученные в распределителях и агрегированные в список пар `[(k2, [v2])]`. Функция-редуктор (`reduce()`) обрабатывает каждую пару списка по отдельности и формирует новую пару `(k3, v3)`, которые собираются в итоговый результат и записываются в блоки HDFS. Количество задач редукции определяется независимо от размеров входных данных. Задачи редукции не пользуются локальностью данных; входные данные одной задачи редукции образуются из выходных данных всех задач распределения: отсортированный вывод распределений передается по сети на узел, где работает задача редукции, на которой данные объединяются и передаются в функции-редукторы. Вывод функций-редукторов хранится в HDFS таким образом, что первая реплика хранится на локальном узле, а другие — на внесегментных узлах.

Схематичное изображение последовательности вычислений MapReduce показано на рис. 2.

За счет того, что MapReduce относится к категории архитектур без разделения (то есть с независимыми задачами), то он защищен от сбоев, возникающих на отдельных машинах кластера, перезапуская неудачные задачи на работоспособных машинах. Этот эффект поддерживается также механизмом репликации в HDFS.

2.2.4. Программный интерфейс Hadoop Streaming API

Hadoop Streaming API [49] — прикладной программный интерфейс для MapReduce, позволяющий использовать для создания функций распределения и отображения языки программирования, отличные от Java. Реализована данная возможность средствами стандартного потока Unix так, что при написании программ MapReduce можно использовать любой язык, взаимодействующий со стандартным потоком ввода и стандартными потоками вывода и ошибок. Функции распределения и редукции читают строки сплита данных из стандартного ввода `STDIN`, которые отсортированы по ключу и записывает результат в стандартный поток вывода `STDOUT`. Данные потоки представлены в языке программирования Python, соответственно, переменной `sys.stdin` и функцией `print()`.

2.3. Облачная платформа Microsoft Azure HDInsight

Microsoft Azure HDInsight [52] — облачный сервис, позволяющий удаленно развертывать кластеры Hadoop и предоставляющий программную среду для операций управления, анализа и визуализации больших данных. HDInsight полностью совместим с открытыми стандартами Apache Hadoop и может обрабатывать неструктурированные или частично структурированные варианты данных. За счет интеграции с платформой данных Hortonworks [46] пользователь имеет возможность перемещать данные из локальных и других облачных источников в облако Azure для создания резервных копий, разработки и тестирования. Встроенная система аналитики позволяет отправлять запросы и к облачным, и к локальным кластерам.

Программные расширения HDInsight доступны для языков C#, Java, .NET, Python, JavaScript.

Максимально доступное количество узлов предоставляемого кластера — 32 узла, причем плата взимается только за реально используемые вычислительные ресурсы. После создания кластера, к нему имеется полный доступ, в том числе по RDP; доступно управление через панель администрирования, которая в том числе позволяет создавать задачи для расчета на кластере. Кроме задач MapReduce, поддерживается интерактивная консоль для составления запросов на JavaScript и Hive.

Магазин данных Azure Marketplace собирает в режиме реального времени данные, изображения и веб-службы от коммерческих поставщиков данных и из официальных общедоступных источников информации. Сервис упрощает процессы приобретения и использования данных: демографических, финансовых, связанных с окружающей средой, розничной торговлей, спортом и пр.

Хранилище HDInsight интегрировано с удаленным хранилищем BLOB-объектов Azure, которое представляет данные в двоичном формате. Так как оно не привязано к конкретному кластеру, хранимые данные можно использовать многократно для разных задач. Компоненты HDInsight напрямую взаимодействуют со структурированными и неструктурированными данными в хранилище больших двоичных объектов через интерфейс HDFS. Структура хранилища данных HDInsight представлена на рис. 3.

В контейнерах BLOB-объектов данные хранятся в виде пар ключ/значение, при этом отсутствует иерархия каталогов. Обращение к файлам в хранилище BLOB-объектов осуществляется через схему URI. Также для этих целей можно использовать Azure CLI (интерфейс командной строки Azure) или Azure PowerShell.

Доступ к удаленному кластеру из-под операционной системы осуществляется посредством SSH, по RSA-ключу, либо по паролю. Поддерживается туннелирование SSH для доступа к веб-интерфейсу Ambari, ResourceManager, JobHistory, NameNode, Oozie и др. Выполнение MapReduce задач на языках, отличных от Java, реализуется посредством упомянутого выше Hadoop Streaming API.

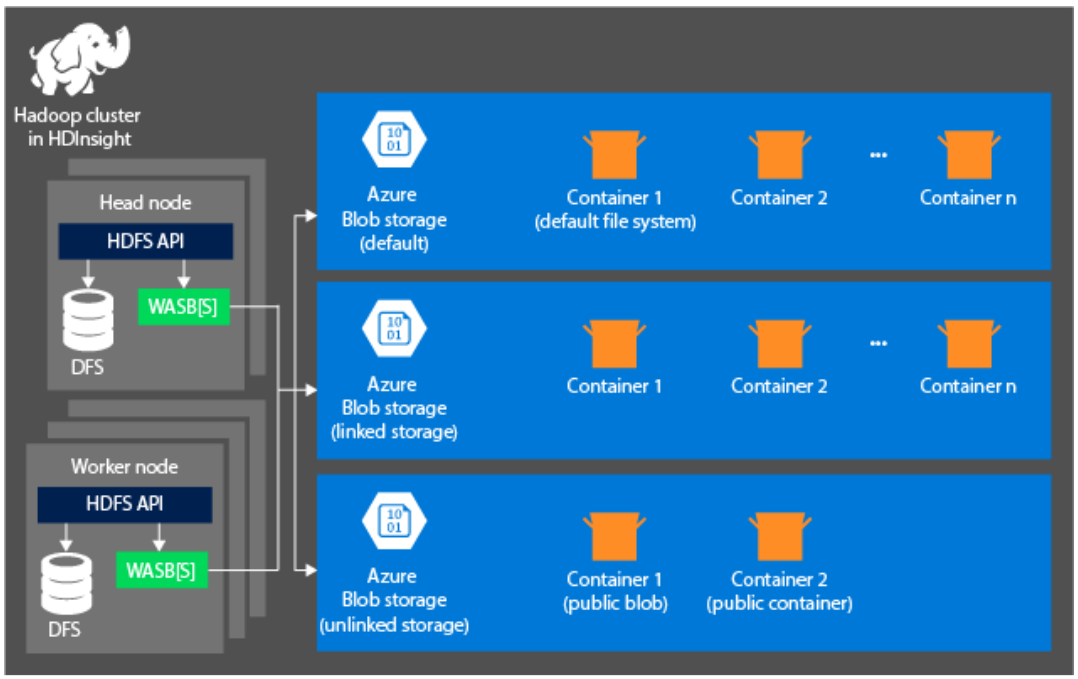

Рисунок 3 — Структура хранилища HDInsight

2.4. Большие данные и исследования общественного транспорта

Области применения больших данных зависят от распространения вычислительных ресурсов, их мощности, увеличения ресурсов для хранения данных и, конечно же, от процесса производства различного цифрового контента.

В сфере управления и изучения городского общественного транспорта применение больших данных позволяет достичь очень высоких результатов в точности и достоверности информации о функционировании системы на любом ее масштабе.

Источники больших данных в городах можно разделить на три категории:

1) направленные — генерируемые с использованием традиционных ручных и полуручных методикам наблюдения,

2) автоматизированные — посредством функционирования цифрового устройства, транзакций и взаимодействия в цифровых сетях, данные, собираемые с датчиков и иных встроенных механизмов,

3) добровольные — собираются самими людьми (взаимодействие в социальных сетях, краудсорсинговые данные, GPS-треки).

Из данной тройки лишь автоматически генерируемые данные являются перспективными с точки зрения предоставления полезных наблюдений о городских системах, в том числе, о транспортных системах [23]. Автоматизированные формы наблюдения могут быть получены, к примеру, заменой бумажных проездных билетов в транспорте на автоматические смарт-карты, внедрением системы распознавания номерных знаков транспортных средств, системы автоматического считывания показателей счетчиков скорости, системы автоматизированных служб мониторинга за состоянием дорожного движения и т. д. Очевидно, что каждое из указанных мероприятий обеспечит городские службы управления огромным количеством данных, в особенности в комбинации друг с другом. По этой причине софтверизированный «умный» город, охваченный разнообразными информационными и сенсорными сетями, потребует тех же способов работы с данными, что и традиционные объекты применения методик больших данных, например, социальные сети.

Особое преимущество внедрения информационных систем больших данных состоит в возможности построения надежного информационного базиса для перспективного планирования и прогнозирования. Использование больших выборок, связывание различных форм данных могут обеспечить более глубокий, целостный и надежный анализ.

Средоточение данных, поступающих ото всех элементов городской среды, в единый банк больших данных дает городским службам полезный информационно-аналитический фундамент, однако накладывает требование единообразия такого анализа, как если бы он касался, например, сферы жилищно-коммунального хозяйства города или сферы общественного транспорта. Другими словами, большие и даже «обычные» данные, получаемые из информационных систем различных сфер городского хозяйства, необходимо обрабатывать универсальным способом, позволяющим легко их комбинировать, если того потребует конкретная исследовательская задача.

К примеру, данные о перемещениях подвижного состава системы общественного транспорта разумно отнести к категории «обычных», так как

их объем редко превышает несколько гигабайт. Однако, проведение комплексного исследования функционирования транспортной системы может потребовать комбинировать их с данными о пассажиропотоке и потоках личного автотранспорта, объем которых, в свою очередь, для случая большого города может измеряться десятками гигабайт и более. Очевидно, наличие единого инструментария обработки/анализа позволит избежать сложностей в таком комбинировании, поэтому в роли такого инструментария целесообразно использовать технологии больших данных как обеспечивающие наилучшую работу с очень большими объемами информации.

Глава III. Реализация задачи

1) Описание шагов расчета параметров эффективности общественного транспорта

   – Рассчитываем параметр неэффективности (в разработке), ранжируем городские районы по данному критерию. Получаем список районов и их значения эфективности с точностью до обратного порядка списка.

2) Окончательные результаты, сравнение результатов методик.

3.1. Описание исходных данных

3.1.1. Транспортные данные

Любая имеющаяся система городского общественного транспорта непосредственно управляется и обслуживается одной или несколькими транспортными компаниями, называемыми компаниями-перевозчиками. Перевозчики могут публиковать собираемые ими транспортные данные, представляющие собой пакеты данных с телекоммуникационного оборудования, установленного на подвижном составе, данных с видеокамер, отчеты автоматизированных систем управления, статистические данные и т. д.

GTFS (General Transit Feed Specification) [51] — открытый и оптимизированный для табличного представления формат для расписания общественного транспорта и сопутствующей геопространственной информации, позволяющий использовать эти данные на картах, в планировщиках маршрутов, либо для исследовательско-аналитических целей [1]. GTFS позволяет создавать стандартизированные пакеты транспортных данных (так называемые *каналы*) и публиковать их в свободном доступе как официальным перевозчикам, так и любым пользователям.

Вся необходимая информация GTFS-каналов транспортной системы конкретного города распределена по 13 CSV-файлам с расширением `.txt` и представляется в них в виде строк, разделенных запятыми, за счет чего она сравнительно легка для программной обработки, так как необходимы лишь

комбинирование нужного количества файлов и фильтрация их содержимого в зависимости от целей аналитической работы. Также такой формат данных удобен для реляционного представления и использования в реляционных базах данных.

Каждый файл пакета отвечает за определенную составляющую транспортной информации, их содержимое должно подчиняться жестко заданной структуре. Первая строка каждого файла содержит имена полей данных. Рассмотрим эти файлы по отдельности (примеры содержимого каждого файла см. в приложении 3.7).

1) `agency.txt` — обязательный файл с указанием информации обо всех компаниях-перевозчиках, предоставляющих данные для данного GTFS-канала.

2) `calendar.txt` — обязательный файл с датами для маршрутов, функционирующих по недельным графикам; описывается начало и конец действия маршрута, а также дни недели, когда маршрут доступен.

3) `calendar_dates.txt` — необязательный файл исключений для маршрутов из `calendar.txt`. Если файл `calendar_dates.txt` включает все даты для маршрута, то этот файл может быть использован вместо файла `calendars.txt`.

4) `fare_attributes.txt` — необязательный файл с тарифами, установленными компаниями-перевозчиками.

5) `fare_rules.txt` — необязательный файл правил действия тарифов, установленных компаниями-перевозчиками. Необязательный файл.

6) `feed_info.txt` — необязательный файл с дополнительной информацией о канале (контактные данный организатора канала, версия, актуальность публикуемой информации).

7) `frequencies.txt` — необязательный файл интервалов между рейсами для маршрутов с переменными интервалами.

8) `routes.txt` — обязательный файл всех маршрутов, обслуживаемых компаниями-перевозчиками.

9) `shapes.txt` — необязательный файл правил отображения траекторий движения маршрутов на карте.

10) `stops.txt` — обязательный файл с информацией о местах, где транспортные средства осуществляют посадку/высадку пассажиров.

11) `stop_times.txt` — обязательный файл с перечислением времени прибытия и отправления транспортного средства в каждом рейсе на каждой остановке.

12) `transfers.txt` — необязательный файл правил осуществления пересадок между маршрутами.

13) `trips.txt` — обязательный файл рейсов, совершаемых каждым маршрутом, представляемые в виде последовательностями остановок.

Диаграмма связей между файлами GTFS-канала представлена на рисунке 4.

Данные, распространяемые в формате GTFS, полностью покрывают нужды данного исследования, следовательно могут использоваться в качестве исходной транспортной информации. В качестве примера были выбраны транспортные данные формата GTFS, опубликованные компанией-перевозчиком TTC города Торонто, Канада в открытом доступе на портале портале GTFS Data Exchange [47]. Дальнейшая часть работы описывает последовательность действий по анализу транспортной системы именно этого города.

Из общего пакета загруженных данных размера 265 Мбайт для целей исследования необходимы 5 файлов: `routes.txt`, `shapes.txt`, `stops.txt`, `stop_times.txt` и `trips.txt`.

3.1.2. Административное деление и население

Деление исследуемого города на районы может выполняться произвольным образом; в рамках данной работы было выбрано официальное

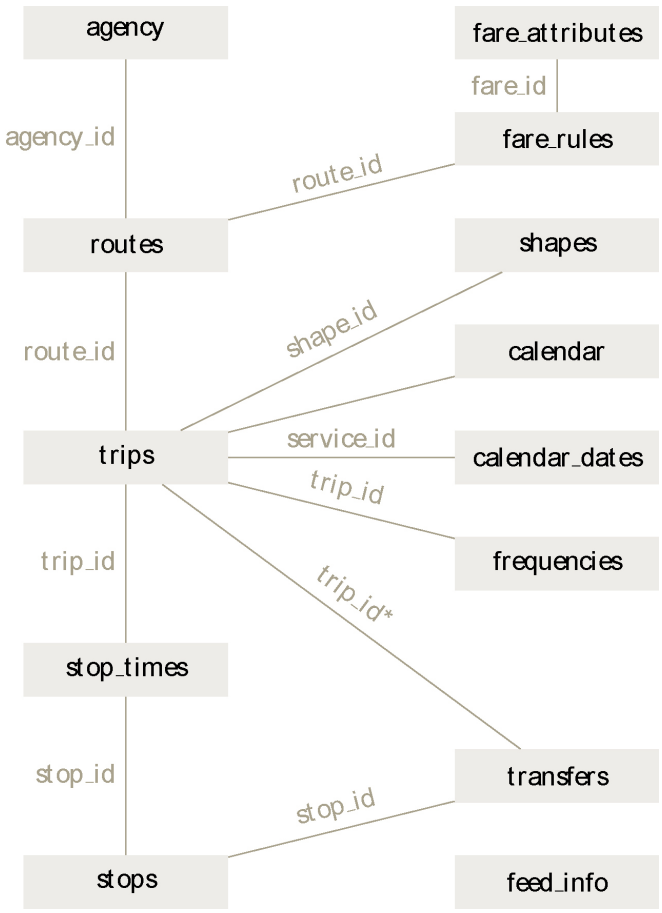

Рисунок 4 — Диаграмма файлов GTFS

административное деление города Торонто. Данные об этом разбиении были загружены в виде векторного shape-файла полигонов с привязкой к координатной проекции EPCG:4326–WGS-84 с портала Открытых данных Торонто [41].

Для получения данных о численности населения с разбивкой по районам на основе административного деления города Торонто необходимо обратиться к источнику [42], содержащему статистические демографические данные.

3.1.3. Улично-дорожная сеть

Необходимый базовый картографический слой требует наличия улично-дорожной сети исследуемого города, привязанной к координатной сетке. Данный слой играет роль пространственного каркаса территории города, к которому привязываются объекты остановок, извлекаемые из пакета транспортных данных, и полигоны районов.

В качестве источника данных об улично-дорожной сети города Торонто использовался картографический сервис OpenStreetMap [48], обладающий свободной лицензией на право использования и загрузки произвольных комбинаций картографических данных и позволяющий вносить в них любые изменения.

Карта OpenStreetMap имеет топологическую структуру данных и хранится в XML-файле с расширением `.gpx`; вся имеющаяся информация сервиса относится к одной из трех базовых сущностей: точкам, линиям/полигонам и отношениям, выражаемым в XML-тегах, соответственно, `<node>`, `<way>` и `<relation>`. Любая дополнительная информация назначается этим примитивам добавлением требуемого количества дочерних тегов `<tag>`, каждый из которых содержит два свойства: имя `k="name"` и значение `v="value"`. В частности, для линий `<way>` это может быть тип дороги (`<tag k="highway"v="residential"/>`), наименование дороги (`<tag k="name"v="Clipstone Street"/>`) и т. п.

Общий вид простейшего файла карты OpenStreetMap, содержащего точку дорожного знака и линию дороги, состоящую из одного звена:

```
<?xml version="1.0" encoding="UTF-8"?>
<osm version="0.6" generator="CGImap 0.0.2">
  <node id="1831881213" changeset="12370172" lat="54.0900666"
  lon="12.2539381" user="lafkor" visible="true">
    <tag k="traffic_sign" v="city_limit"/>
  </node>
  <way id="26659127" user="Masch" uid="55988" visible="true"
  changeset="4142606">
    <nd ref="292403538"/>
    <nd ref="298884289"/>
    <tag k="highway" v="unclassified"/>
  </way>
</osm>
```

Стандартный интерфейс загрузки данных с основных серверов OpenStreetMap предполагает скачивание полного файла карты выбранной

области, содержащего все находящиеся на ней объекты. Для выборочного (и более быстрого) извлечения из общей карты только необходимых категорий объектов был использован прикладной веб-интерфейс Overpass API [45], позволяющий вместо загрузки полного файла карты направлять конкретные запросы выделенным серверам в обход основных. Интерфейс имеет собственный язык запросов Overpass QL, который, в свою очередь, обладает унифицированной программной оберткой на языке программирования Python [50].

3.2. Программное обеспечение и конфигурация

Исследование частично проводилось в рамках свободно распространяемой геоинформационной системы QGIS [44]. Инструментальный пакет данной системы позволил упростить действия по присоединению разнородных столбцовых транспортных и векторных картографических данных на основе пространственных таблиц PostGIS, сведя к минимуму их ручное связывание и обработку, дополнительно предоставляя возможность реализации отсутствующего программного функционала самописными модулями на языках программирования Python и C++ за счет наличия программного API QGIS. Также был задействован следующий инструментарий:

- Пространственный анализ
- Пространственные запросы
- Функции буферизации
- Наложение полигонов
- Оценка близости и определение пересечения улиц
- Представление полученных результатов

Для обработки транспортных данных на портале Microsoft Azure был развернут облачный Hadoop-кластер HDInsight с предустановленным программным пакетом, необходимым для функционирования экосистемы. Конфигурация кластера:

1) 2 головных узла с 4-ядерными центральными процессорами и по 14 ГБ ОЗУ каждый. Также головные узлы оснащаются локальными дисками SSD объемами 200 ГБ.

2) 4 рабочих узла с 4-ядерными центральными процессорами и по 7 ГБ ОЗУ каждый.

3) Общий объем диского хранилища, выделенного под данный кластер — 8 ТБ.

Базовая операционная система кластера — Ubuntu Linux 12.04 LTS. Доступ к головному узлу кластера осуществляется удаленно посредством SSH с аутентификацией по RSA-ключу.

3.3. Подготовка данных

3.3.1. Конвертация файла УДС

В качестве первого шага выполним конвертацию файла улично-дорожной сети, имеющего, как уже было сказано, формат XML с расширением `.gpx`, в векторный shape-формат, требуемый для дальнейшего использования в геоинформационной системе. Данное действие было произведено с использованием языка программирования Python и библиотечных модулей языка `ElementTree` и `pyshp`.

3.3.2. Выгрузка данных в HDFS облачного кластера

Для работы с транспортными данными их требуется выгрузить в облачное хранилище, где будут производиться основные действия по обработке. С этой целью необходимо получить удаленный доступ к одному из головных узлов кластера и выполнить внутреннюю команду `hadoop -copyFromLocal <localFilePath> <storageFilePath>`. Стоит отметить, что файловая система HDFS кластера по умолчанию находится в хранилище BLOB-объектов Azure, поэтому для доступа к директориям и файлам, хранящимся в HDFS используется следующий синтаксис URI: `wasb[s]://<container>@<account>.blob.core.windows.net/<path>`,

либо его укороченная версия, включающая непосредственно протокол `wasb:` и путь в дереве файловой системы.

3.3.3. Фильтрация данных

Далее выгруженные файлы с данными отфильтровываются от лишних данных. Сохраняемые поля файлов:

1) Файл `stops.txt`: `stop_id`, `stop_name`, `stop_lat`, `stop_lon`.

2) Файл `routes.txt`: `route_id`, `route_short_name`, `route_type`.

3) Файл `trips.txt`: `route_id`, `trip_id`, `direction_id`, `shape_id`.

4) Файл `stop_times.txt`: `trip_id`, `arrival_time`, `stop_id`, `stop_sequence`.

5) Файл `shapes.txt`: `shape_id`, `shape_pt_lat`, `shape_pt_lon`, `shape_pt_sequence`.

Также содержмое файла `stop_times` фильтруется по временному промежутку, если это требуется, например, для исследования за определенное время суток.

Фильтрация осуществляется с помощью высокоуровневой программной платформы Apache Pig, предустановленной в HDInsight и ее процедурного языка Pig Latin. Доступ к командной строке-интерпретатору grunt производится вводом команды `pig` в консоли одного из головных узлов удаленного кластера. Программы Pig Latin следуют общему шаблону:

1) Загрузка: чтение данных для обработки из файловой системы с помощью команды `LOAD 'wasb:///path/' as (fields:type); DUMP <data>;`. Результатом работы оператора `LOAD` является отношение, представляемое набором кортежей, которые состоят из полей в строках по данным входного файла.

2) Преобразование: в данном случае пример фильтрации файла `stop_times` с помощью команды `<filtered_data> = FOREACH`

<data> GENERATE $0, $1, $3, $4;, которая составляет структуру полей, состоящих из полей по их индексам в старой структуре.

3) Дамп или сохранение: сохранение полученного файла для дальнейшей обработки с помощью команды STORE RESULT into 'wasb:///path/'; QUIT.

3.3.4. Выполнение задания MapReduce

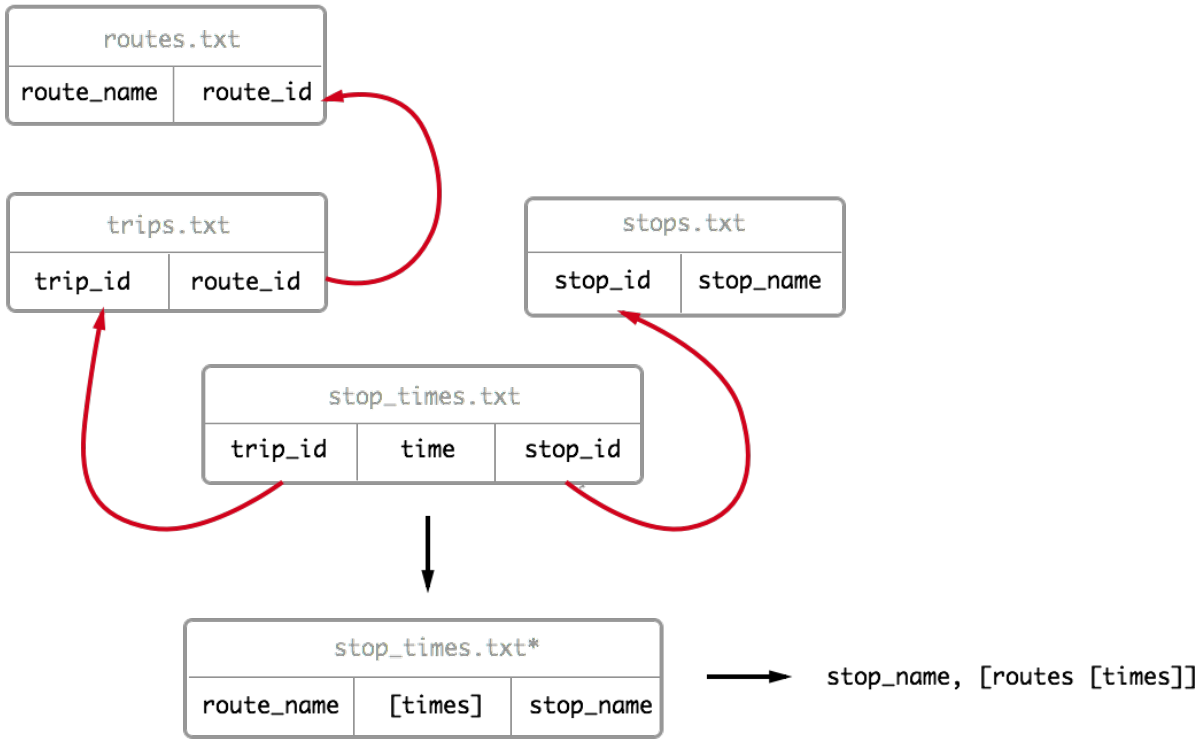

Рисунок 5 — Обобщенная схема преобразования транспортных данных

Отфильтрованные данные требуется скомбинировать таким образом, чтобы получить ранжирование остановок по частоте рейсов на каждой из них. Для этого необходимо создать файл, каждая строка которого будет содержать идентификатор остановки, ее название и список пар всех маршрутов, которые обслуживает данная остановка и списков рейсов, совершенных в рамках соответствующего маршрута. Общий вид строк данного файла:

```
stop_id, stop_name, stop_lat, stop_lon,
[(route_short_name, [arrival_times])], times_num
```

Так как в комбинации участвуют несколько файлов с данными, то необходимо произвести их полное внешнее соединение до того, как они попадают в функцию распределения. Файлы `routes` и `trips` соединяются по ключу поля `route_id`; файл `stop_times` соединяется с файлом `stops` по ключу поля `stop_id` и с файлом-объединением, полученным раннее, по ключу поля `trip_id`. Для выполнения соединения на стороне отображения используется пакет `org.apache.hadoop.mapreduce.join`. Источники ввода и тип соединения настраиваются через выражение соединения, написанное по простой схеме. Общий вид соединений и преобразований данных показан на рис. 5.

В функцию распределения полученная структура загружается построчно с ключом значения `sys.stdin` (соединенного файла) и значением содержимого этого файла. Выход функции содержит список строк файла без полей `trip_id` и `route_id` следующего вида:

```
stop_id, stop_name, stop_lat, stop_lon,
(route_short_name, arrival_time), 1
```

Далее производятся три операции комбинации, в ходе первой из которых все единицы с одинаковыми значениями поля `stop_id` агрегируются в список, в ходе второй комбинации все значения поля `arrival_time` с одинаковыми значениями поля `route_short_name` агрегируются в список, а в ходе третьей комбинации в список агрегируются все пары `(route_short_name, [arrival_times])` с одинаковыми значениями поля `stop_id`.

В функцию редукции подается список полученных строк следующего вида:

```
stop_id, stop_name, stop_lat, stop_lon,
[(route_short_name, [arrival_times])], [ones]
```

Выход функции редукции содержит файл с данными строками, в которых списки единиц были собраны в итоговое поле общего количества совершенных рейсов на каждой остановке.

Для составленного кода MapReduce в консоли одного из головных узлов удаленного кластера вызывается соответствующая задача с помощью команды `yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-`

`streaming.jar -files mapper.py,reducer.py -mapper mapper.py -reducer reducer.py -input wasb:///input/path/ -output wasb:// /output/path/`.

Эта команда состоит из следующих частей:

- `hadoop streaming.jar` — используется при выполнении операций потоковой передачи MapReduce. Эта часть обеспечивает взаимодействие Hadoop с предоставленным внешним кодом MapReduce.
- `-files` — сообщает Hadoop, что для этого задания MapReduce необходимы указанные файлы, которые нужно скопировать на все рабочие узлы.
- `-mapper` — сообщает Hadoop о том, какой файл следует использовать в качестве модуля распределения.
- `-reducer` — сообщает Hadoop о том, какой файл следует использовать в качестве модуля редукции.
- `-input`, `-output` — входной файл и выходной каталог.

Для загрузки файла с результатами из облачного BLOB-хранилища кластера составляется сценарий, приведенный в приложении 3.7

3.3.5. Перемещение данных в геоинформационную систему

Все данные, находящиеся в геоинформационной системе являются набором слоев, которые должны иметь векторный формат. Каждый слой представляется таблицей атрибутов, где каждая строка описывает отдельный объект слоя заданным набором атрибутов-свойств, описываемых множеством столбцов.

Тогда общий набор слоев в геоинфорационной системе состоит из: векторного представления улично-дорожной сети, векторного представления административного деления, таблицы с данными о населении, файла ранжированных остановок.

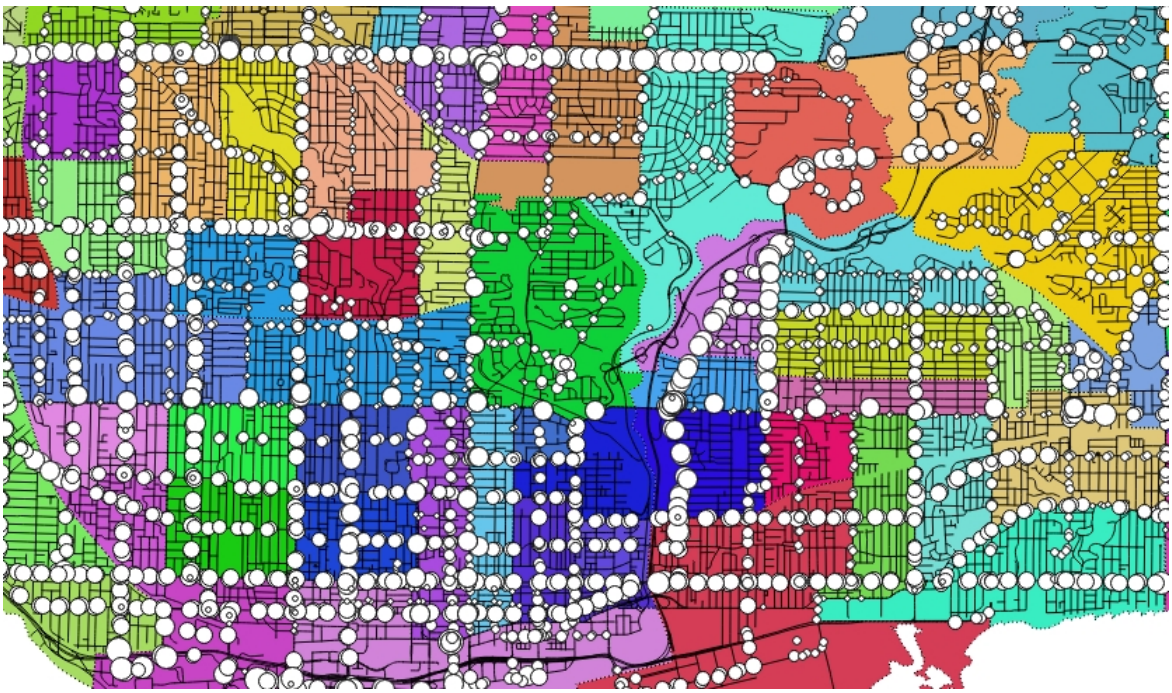

Рисунок 6 — Визуальное представление набора слоев данных в геоинформационной системе

Улично-дорожная сеть привязывается к полигонам районов путем установления географических координат точек дорог на территории соответствующих вмещающих полигонах. Также к слою административного деления присоединяется таблица численности населения путем соединения двух таблиц атрибутов по ключу названия района.

Слой остановок является точечным, на основании установления атрибутов координат в качестве географических. Точки остановок присоединяются к слою улично-дорожной сети аналогично вышеописанному присоединению. Дополнительно все остановки сортируются по атрибуту количества рейсов, что сказывается на их визуализации: чем больше рейсов совершено на остановке, тем крупнее ее маркер точки.

Визуализация полученного набора слоев показана на рис. 6.

3.4. Реализация анализа параметра доступа

Напомним, что расчет данного параметра заключается в выясении степени покрытия территории заданного района буферными зонами остановочных пунктов, обсуждавшихся в разделе 1.3.1.

В геоинформационной системе реализация анализа для каждого района исследуемого города производится с использованием инструментария пространственного анализа с применением буферов. Его идея заключается в создании буферов объектов присоединяемого слоя, в данном случае слоя остановок и их наложения на полигоны основного слоя, в данном случае слой административного деления с использованием заданной функции.

В соответствии с рассуждениями в разделе 1.3.1, зададим радиус буферной зоны для точек остановок равным 400 метрам. Далее производится расчет процентного значения покрытия территории района буферами находящихся на ней остановок. Для этого реализуется пространственный анализ буферов остановок и полигонов районов административного деления с использованием функции пространственной разности; непокрытые буферами фрагменты полигонов районов выделяются в отдельный векторный слой «необслуживаемых» территорий.

На следующем шаге производится вычисление площади каждого полигона района и запись полученных значений в таблицу слоя административного деления как нового атрибута. Точно так же вычисляются площади каждого полигона слоя «необслуживаемых» территорий, которые в свою очередь дополнительно суммируются, если несколько полигонов являются фрагментами одного района.

Наконец, производится связывание таблиц слоя районов и слоя необслуживаемых территорий, вычисляется процентное отношение необслуживаемой площади к полной, значение которого присоединяется к таблице слоя административного деления в виде дополнительного атрибута. Фрагмент итоговой преобразованной таблицы слоя административного деления представлен в табл. 1.

Визуализация результатов произведенного пространственного анализа имеют вид, показанный на рис. 7. Легенда обозначает процентные отношения площади «необслуживаемых» территорий к общим площадям районов.

Заданная стандартная дистанция доступа позволяет объективно проанализировать исследуемую характеристику охвата; это можно проиллюстрировать, рассмотрев другие различные значения дистанции, к примеру, от 200

AREA_CD	AREA_NAME	AREA	NO_TRANSIT	PERCENT
097	York Un. ...	1685997	2082	1
027	Yonge-St. ...	19247065	1340410	7
038	Lansing ...	7764098	3309051	43
031	Yorkdale ...	8770819	163855	2
...				

Таблица 1 — Таблица слоя административного деления в результате расчета покрытия. Первый столбец — идентификатор района, третий и четвертый столбцы — соответственно, общая площадь района в м^2 и площадь «необслуживаемой» территории

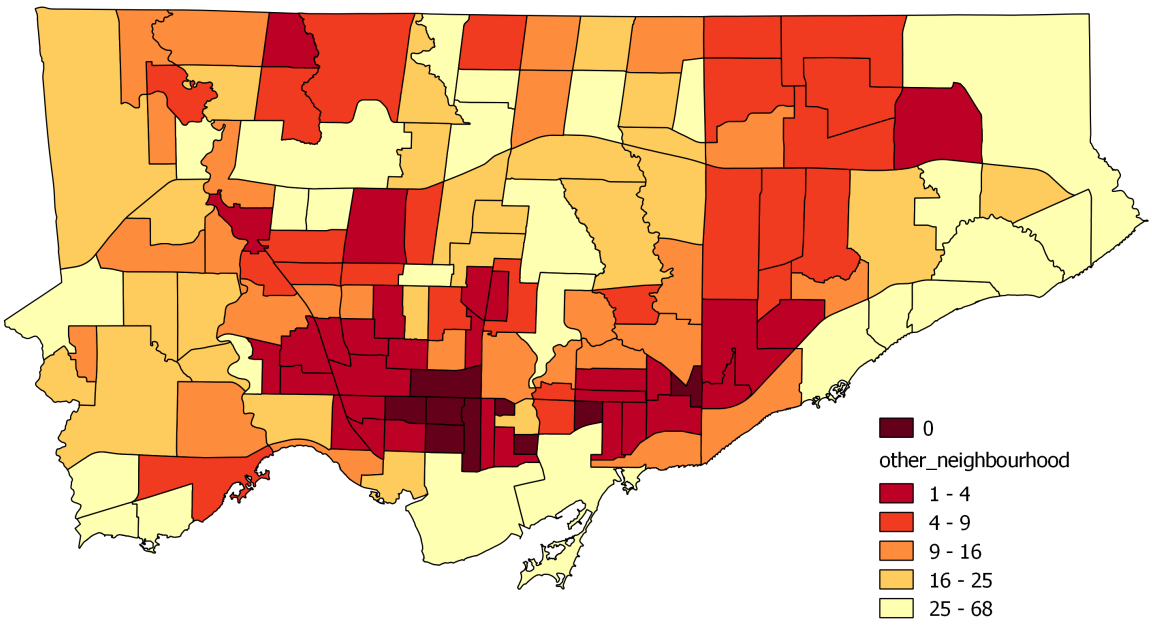

Рисунок 7 — Визуализация результатов вычисления значений покрытия. Разбивка по цветам означает степень покрытия

до 600 метров и получив соответствующие значения процента охвата населения, что отображено на графике 8. Видно, что вблизи значения в 400 метров отсутствуют скачки графика, что позволяет утверждать о разумности его выбора в качестве стандартной величины. Логарифмическая зависимость между покрытием населения и дистанцией удовлетворительного доступа к остановкам означает, что население сосредоточено в определенных областях так, что для доступа к более дальней остановке вместо ближайшей требуется преодолеть значительное расстояние [29].

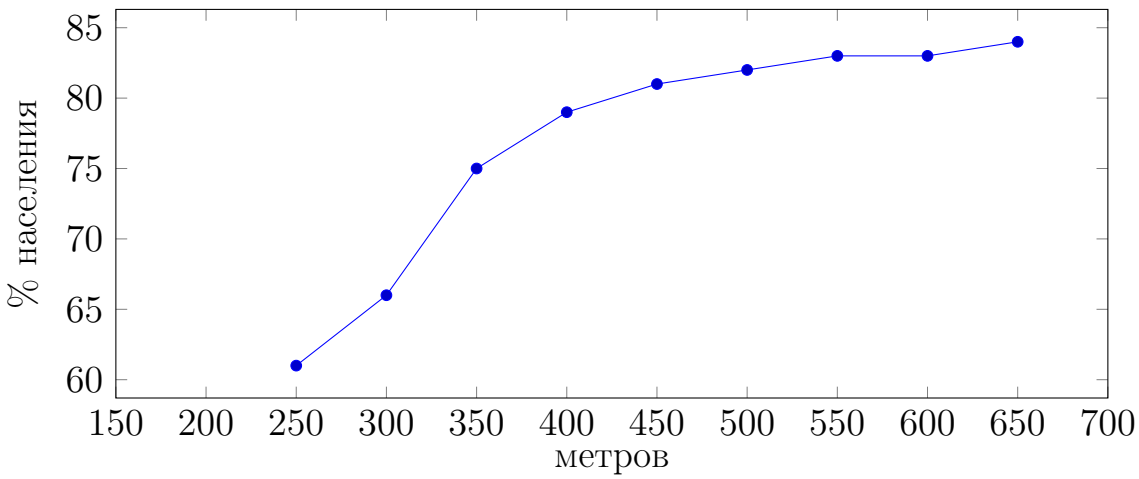

Рисунок 8 — График отношения радиуса буферной зоны удовлетворительной доступности остановок и соответствующего ему процента охватываемого населения

3.5. Реализация анализа параметра доступности

Аналогично предыдущему подразделу напомним, что расчет данного параметра для отдельного района заключается в нахождении среднего значения от числа совершаемых рейсов на всех остановках, расположенных в пределах этого района.

В геоинформационной системе для выполнения такого расчета применен инструментарий пространственного запроса, основанный на анализе вхождения объектов одного слоя (в данном случае точек остановок) в границы другого слоя, обязательного представленного полигонами (в данном случае полигонами слоя административного деления), и манипулирования значениями какого-либо атрибута всех вошедших в пределы полигона объектов. Так как слой остановок ранее получил ранжирование по количеству совершенных на каждой остановке рейсов, то в процессе пространственного запроса будем агрегироваться по данному параметру. Механизм пространственного запроса позволяет указать математическую функцию, применяемую к агрегированному значению: в качестве нее зададим медианную функцию.

Полученное медианное значение присоединяется каждому объекту района в таблице слоя административного деления как новый атрибут, после чего она приобретает вид, частично показанный в табл. 2.

AREA_CD	AREA_NAME	TRIPS_COUNT
097	York Un. ...	156
027	Yonge-St. ...	111
038	Lansing ...	669
031	Yorkdale ...	709
...		

Таблица 2 — Таблица слоя административного деления в результате расчета частоты

Визуализация результатов произведенного пространственного анализа имеет вид, показанный на рис. 9. Легенда обозначает значения параметра средней частоты рейсов.

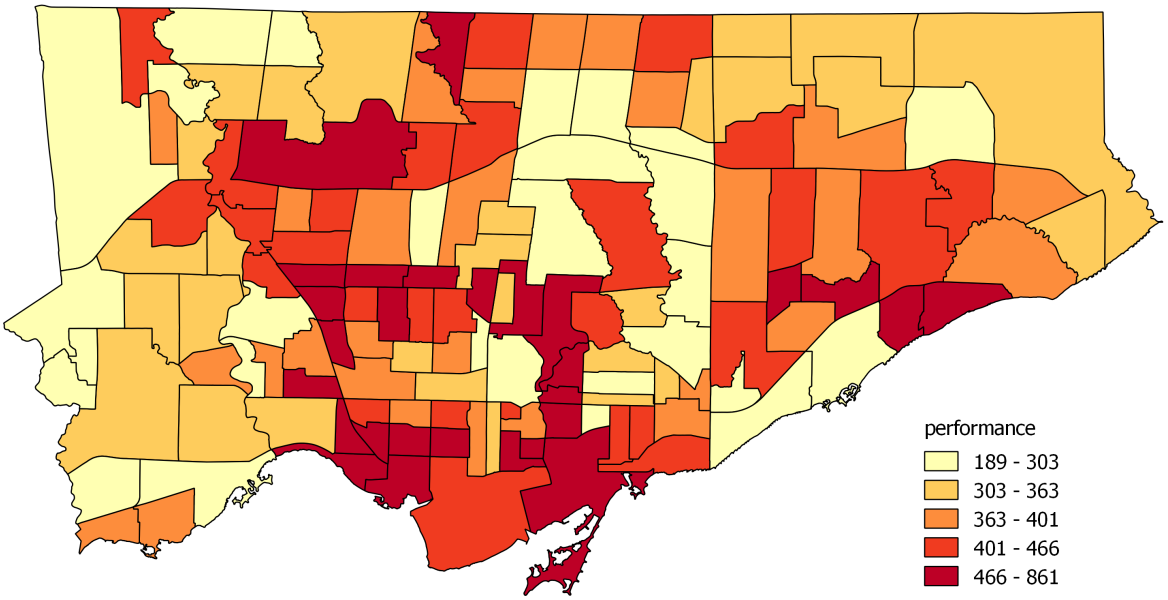

Рисунок 9 — Визуализация результатов вычисления значений частоты рейсов. Разбивка по цветам означает среднее количество рейсов

3.6. Совмещенный анализ параметров доступа и доступности

Сравнение результатов, полученных в предыдущих двух подразделах, позволяет сделать предположение о необходимости более общей и объективной оценки: в случае города Торонто не все районы с наилучшим покрытием маршрутной сетью общественного транспорта (от 91% покрытия площади

территории), являются лидерами по средней частоте совершаемых рейсов (400 и более). Это наблюдение сигнализирует о важности учета обоих параметров при изучении эффективности функционирования общественного транспорта: одно лишь максимальное покрытие маршрутной сетью территории района не всегда влечет удовлетворительную эффективность общественного транспорта в его пределах.

Принимая во внимание это наблюдение, необходимо совместить оба модифицированных слоя административного деления, полученных выше, в один с новым атрибутом, который совмещает атрибуты со значением покрытия и со значением частоты с применением набора логических условий. Полученный новый атрибут характеризует, таким образом оба параметра и позволяет определять район как эффективный, если на его территории находится большое количество остановок с большим количеством частых маршрутов, и наоборот, как неэффективный, если на его территории расположено недостаточное количество остановок, либо большинство из них обслуживает малое количество маршрутов или маршруты с редким движением. Значения атрибута выражены в процентах.

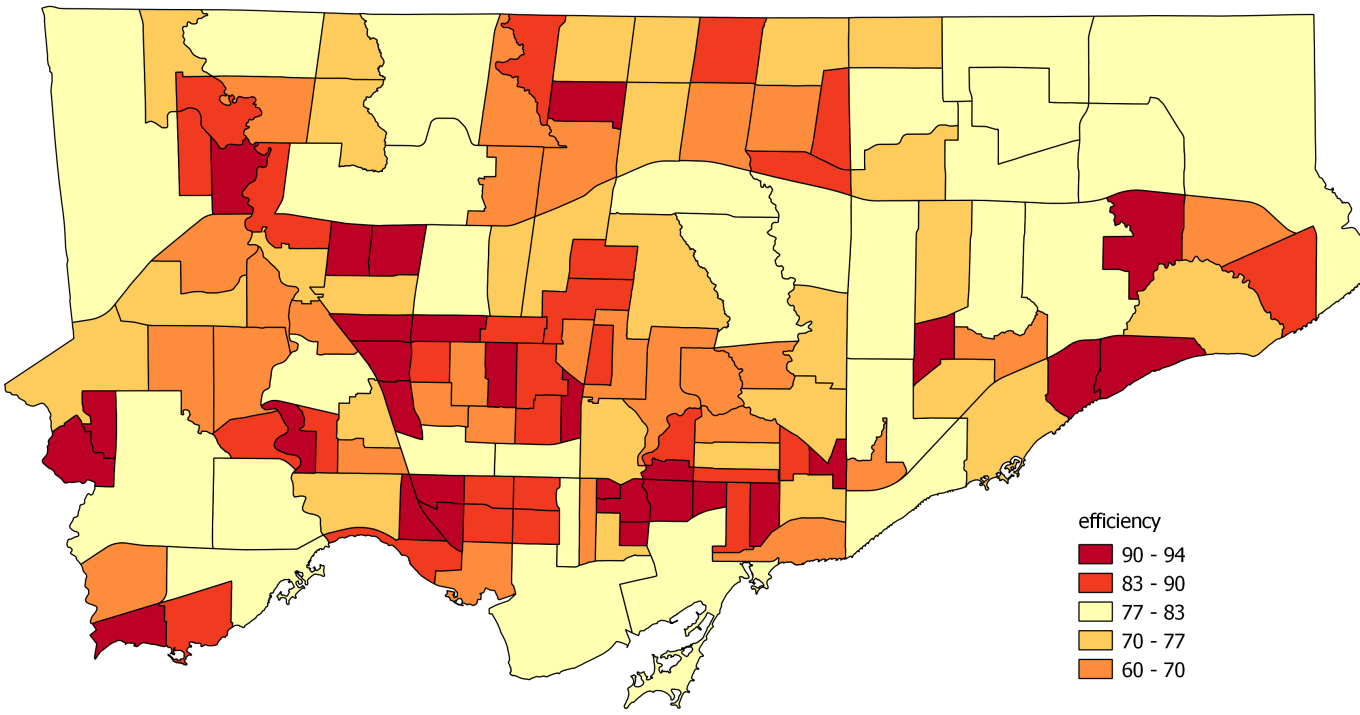

Рисунок 10 — Визуализация совмещения параметров покрытия и частоты. Разбивка по цветам характеризует процентное значение совмещенного атрибута эффективности

Визуализация результатов полученного совмещенного параметра эффективности показанна на рис. 10.

3.7. Реализация анализа избыточности

Как было описано в разделе 1.5.3, избыточность размещения остановок общественного транспорта может являться характеристикой неэффективности функционирования транспортной системы, и соответствующая задача (1) имеет решение, реализуемое Лагранжевой эвристкой, с помощью которой производится нахождение наименьшей стоимости покрытия множества (в данном случае, территории города остановочным пунктами).

Решение данной эвристики определяется нижеследующим алгоритмом 1 (столбцово-ориентированная процедура сокращения (column-domination reducing) описана в [3]).

Algorithm 1 Алгоритм решения Лагранжевой эвристики

```
1:  procedure LHSCP
2:      UB ← ∞
3:      Выполнить процедуру column_domination_reducing
4:      for all t субградиентного метода do
5:          Solve R(π^(t))
6:          optimal ← x^(t)
7:          Проверить, является ли optimal покрытием
8:          if не является then
9:              завершить с использованием жадной эвристики
10:         end if
11:         Наименьшая извлекаемая стоимость текущего покрытия ← (7)
12:         UB ← update UB
13:         π^(t) ← update π^(t)
14:     end for
15: end procedure
```

Выходная переменная UB будет являться искомой наименьшей стоимостью покрытия. Сложность решения $R(\pi)$, проверки оптимального решения

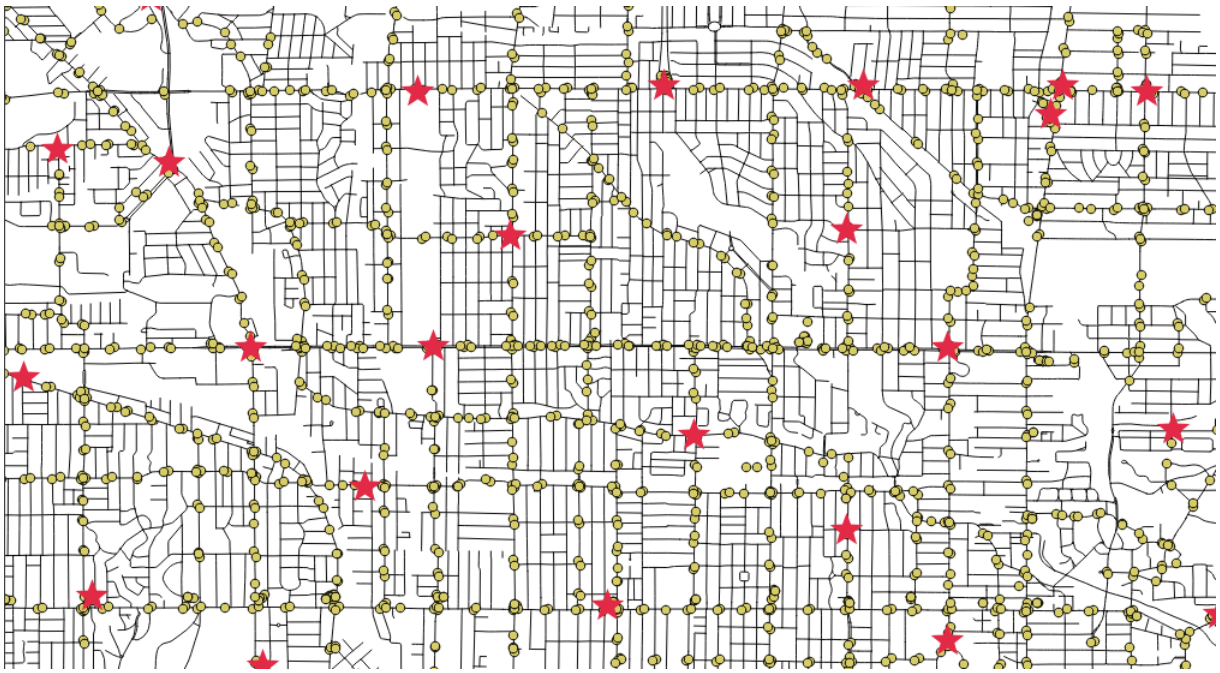

Рисунок 11 — Визуализация результатов поиска избыточных остановок

на покрытие, обновления вектора π составляет $O(f)$, где $f = \sum_{j \in J} |P_j|$ [18]. Число итераций алгоритма фиксировано, поэтому общая сложность алгоритма равна $O(m \cdot f)$ по причине доминирования жадной эвристики.

Применение описанного алгоритма к исследуемой транспортной системе позволяет выяснить, какое число остановочных пунктов способно покрывать в смысле доступности такое же количество территории города или такой же процент населения. Программная реализация выполнена на языке программирования Python и запускалась в рамках среды интерпретации, встроенной непосредственно в геоинформационную систему. Обработка, время работы алгоритма на имеющихся даных и отображения результатов потребовали менее одной минуты, однако предельное число итераций не превысило 200 по причине большого масштаба задачи; в связи с этим точность решения колеблется от 0.3% до 5.6%, что вполне приемлемо для данного анализа. Исходное количество остановочных пунктов для города Торонто равно 10693 единицам, по результатам работы алгоритма обнаружено, что из этого числа с сохранением текущего уровня покрытия остается лишь менее 1500 остановок, обеспечивающих исходную эффективность; для соблюдения условия размещения остановок по обе стороны от каждой улицы, данное число эффективных остановок требуется удвоить, получая

72% потенциально избыточных остановок. В итоге, имеется значительный потенциал увеличения скорости передвижения транспортных средств, а следовательно, уменьшения общего времени поездки.

На рис. 11 изображен фрагмент центральной части Торонто с точками исходных остановочных пунктов и звездами сохраняемых эффективных остановок, на рис. 12 — отношение различных буферных расстояний и получаемого количества сохраняемых остановок.

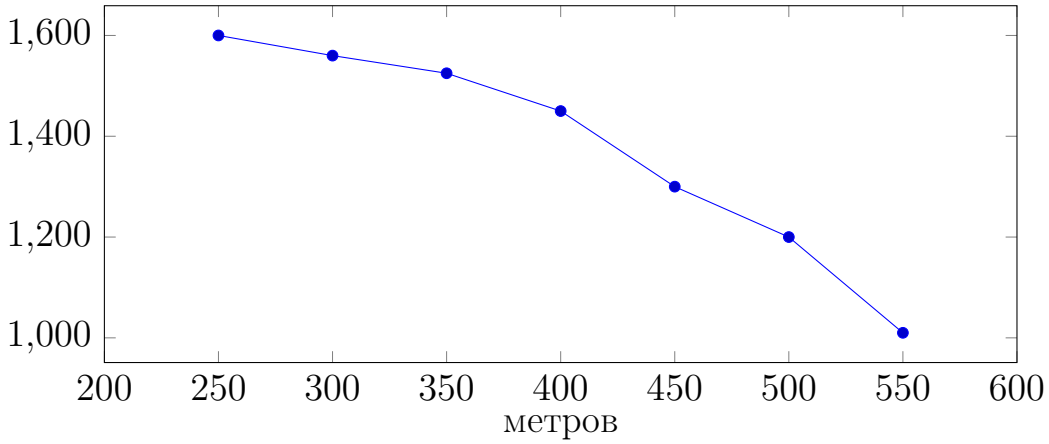

Рисунок 12 — График отношения радиуса буферной зоны удовлетворительной доступности остановок и соответствующего ему количества сохраняемых эффективных остановок

В качестве возможного дополнительного шага целесообразно ввести различные радиусы буферных зон доступности остановок, например, уменьшить его до 100 метров в центральной, наиболее застроенной части исследуемого города и увеличить до 500 метров в периферийных районах, дополнительно учитывая веса самих остановок, которые отражали бы их общегородскую важность. Также на практике может потребоваться закреплять какие-либо важные по разным причинами остановочные пункты так, чтобы ни один из них не был удален, если его расположение окажется избыточным. К примеру в крупных пересадочных узлах нельзя удалять остановки, входящие в них, даже если их расположение не отвечает требованиям к эффективности.

Выводы

В данной работе была представлена последовательность действий по реализации анализа эффективности работы общественного транспорта на примере города Торонто.

Для выполнения исследования прежде всего требуется обратить особое внимание на выбор параметров эффективности в виду их многообразия и слабой формализуемости.

Продемонстрирована применимость методов анализа больших данных в исследовательских целях, конкретно при изучении функционирования транспортной системы.

Однако приведенный способ имеет возможности для усовершенствования, таких как:

– Добавление дополнительного ранжирования объектов улично-дорожной в пределах района города так, что ранг конкретной улицы мог бы зависеть от ее важности и пропускной способности. Тогла это свойство выступает дополнительной характеристикой остановок: например, если остановка расположена на важной магистрали, то высокий ранг ей обеспечит большее количество маршрутов/рейсов, нежели в случае ее расположения на второстепенной улице.

– Добавление характеристик численности населения районов города для ответа на вопрос, какое значение транспортной насыщенности будет удовлетворительным для конкретного района.

– Добавление информации о пассажиропотоке, которую можно получать с данных электронных проездных билетов (смарт-карт). Исследование пассажиропотока и передвижения ОТ сообща могло бы дать обширные знания о востребованности пассажирами тех или иных маршрутов.

– Для получения релевантной статистики имеет смысл получить данные за больший, чем суточный, временной промежуток — к примеру, движение транспорта значительно различается в будние, выходные и праздничные дни.

Заключение

В рамках проведенного исследования получены следующие результаты:

– Внедрение методики обработки больших данных как универсального аналитического фундамента, пригодного для привязки дополнительных данных, таких как пассажиропоток, потоки автомобильного транспорта и др.

– Анализ результатов вычисления различных параметров эффективности и совмещение некоторых из них для получения более адекватных значений.

Сформулированные задачи были выполнены в полном объеме. Поставленные цели были достигнуты.

Приложение

Пример содержимого GTFS-канала

Файл `agency.txt`:

```
agency_id, agency_name,agency_url,agency_timezone,agency_phone,
agency_lang
FunBus,The Fun Bus,http://www.thefunbus.org,America/Los_Angeles,
(310) 555-0222,en
```

Файл `calendar.txt`:

```
service_id,monday,tuesday,wednesday,thursday,friday,saturday,
sunday,start_date,end_date
WE,0,0,0,0,0,1,1,20060701,20060731
WD,1,1,1,1,1,0,0,20060701,20060731
```

Файл `calendar_dates.txt`:

```
service_id,date,exception_type
WD,20060703,2
WE,20060703,1
WD,20060704,2
WE,20060704,1
```

Файл `fare_attributes.txt`:

```
fare_id,price,currency_type,payment_method,transfers,
transfer_duration
1,0.00,USD,0,0,0
2,0.50,USD,0,0,0
3,1.50,USD,0,0,0
4,2.00,USD,0,0,0
5,2.50,USD,0,0,0
```

Файл `fare_rules.txt`:

```
fare_id,route_id,origin_id,destination_id,contains_id
a,TSW,1,1,
a,TSE,1,1,
a,GRT,1,1,
a,GRJ,1,1,
a,SVJ,1,1,
a,JSV,1,1,
a,GRT,2,4,
a,GRJ,4,2,
b,GRT,3,3,
c,GRT,,,6
```

Файл `feed_info.txt`

```
feed_publisher_name,feed_publisher_url,feed_lang
"СПб ГКУ ""Организатор перевозок""",http://orgp.spb.ru/,ru
```

Файл `frequencies.txt`:

```
trip_id,start_time,end_time,headway_secs
AWE1,05:30:00,06:30:00,300
AWE1,06:30:00,20:30:00,180
AWE1,20:30:00,28:00:00,420
```

Файл `routes.txt`:

```
route_id,route_short_name,route_long_name,route_desc,
route_type
A,17,Mission,"The ""A"" route travels from lower Mission to
Downtown.",3
```

Файл `shapes.txt`:

```
shape_id,shape_pt_lat,shape_pt_lon,shape_pt_sequence,
shape_dist_traveled
A_shp,37.61956,-122.48161,1,0
A_shp,37.64430,-122.41070,2,6.8310
A_shp,37.65863,-122.30839,3,15.8765
```

Файл stops.txt:

```
stop_id,stop_name,stop_desc,stop_lat,stop_lon,stop_url,
location_type,parent_station
S1,Mission St. & Silver Ave.,The stop is located
at the southwest corner of the intersection.,37.728631,
-122.431282,,,
S2,Mission St. & Cortland Ave.,The stop is located
20 feet south of Mission St.,37.74103,-122.422482,,,
S3,Mission St. & 24th St.,The stop is located at the
southwest corner of intersection.,37.75223,-122.418581,,,
S4,Mission St. & 21st St.,The stop is located at the
northwest corner of intersection.,37.75713,-122.418982,,,
S5,Mission St. & 18th St.,The stop is located 25 feet west
of 18th St.,37.761829,-122.419382,,,
S6,Mission St. & 15th St.,The stop is located 10 feet
north of Mission St.,37.766629,-122.419782,,,
S7,24th St. Mission Station,,37.752240,-122.418450,,,
```

Файл stop_times.txt:

```
trip_id,arrival_time,departure_time,stop_id,stop_sequence,
pickup_type,drop_off_type
AWE1,0:06:10,0:06:10,S1,1,0,0,0
AWE1,,,S2,2,0,1,3
AWE1,0:06:20,0:06:30,S3,3,0,0,0
AWE1,,,S5,4,0,0,0
AWE1,0:06:45,0:06:45,S6,5,0,0,0
AWD1,0:06:10,0:06:10,S1,1,0,0,0
AWD1,,,S2,2,0,0,0
AWD1,0:06:20,0:06:20,S3,3,0,0,0
AWD1,,,S4,4,0,0,0
AWD1,,,S5,5,0,0,0
AWD1,0:06:45,0:06:45,S6,6,0,0,0
```

Файл transfers.txt:

```
from_stop_id,to_stop_id,transfer_type,min_transfer_time
S6,S7,2,300
S7,S6,3,
S23,S7,1,
```

Файл `trips.txt`:

```
route_id,service_id,trip_id,trip_headsign,block_id
A,WE,AWE1,Downtown,1
A,WE,AWE2,Downtown,2
```

Скрипт загрузки данных улично-дорожной сети

```
import overpass
api = overpass.API(timeout=600)
output = open('streets.osm', 'w', encoding='utf8')
response = api.Get('''
    way["highway"="motorway"](43.5701,-79.7082,43.8544,
    -79.2275);
    way["highway"="motorway_link"](43.5701,-79.7082,
    43.8544,-79.2275);
    way["highway"="primary"](43.5701,-79.7082,43.8544,
    -79.2275);
    way["highway"="secondary"](43.5701,-79.7082,43.8544,
    -79.2275);
    way["highway"="tertiary"](43.5701,-79.7082,43.8544,
    -79.2275);
    way["highway"="unclassified"](43.5701,-79.7082,
    43.8544,-79.2275);
    way["highway"="residential"](43.5701,-79.7082,43.8544,
    -79.2275)''',
    responseformat=xml)
output.write(response)
output.close()
```

Сценарий загрузки файлов из облачного BLOB-хранилища

```
$resourceGroupName = "<AzureResourceGroupName>"
$clusterName = "<HDInsightClusterName>"
$blob = "path/to/output-file"

$cluster = Get-AzureRmHDInsightCluster -ResourceGroupName
    $resourceGroupName -ClusterName $clusterName
$defaultStorageAccount = $cluster.DefaultStorageAccount
    -replace '.blob.core.windows.net'
$defaultStorageAccountKey = Get-AzureRmStorageAccountKey
    -ResourceGroupName $resourceGroupName -Name
    $defaultStorageAccount |  %{ $_.Key1 }
$defaultStorageContainer = $cluster.DefaultStorageContainer
$storageContext = New-AzureStorageContext -StorageAccountName
    $defaultStorageAccount -StorageAccountKey
    $defaultStorageAccountKey

Write-Host "Download the blob ..." -ForegroundColor Green
Get-AzureStorageBlobContent -Container $defaultStorageContainer
    -Blob $blob -Context $storageContext -Force
```

Список литературы

1. Antrim A., Barbeau S. J. The many uses of GTFS data–opening the door to transit and multimodal applications // Location-Aware Information Systems Laboratory at the University of South Florida. — 2013.
2. Batty M. Big data, smart cities and city planning // Dialogues in Human Geography. — 2013. — Т. 3, № 3. — С. 274—279.
3. Beasley J. E., Jörnsten K. Enhancing an algorithm for set covering problems // European Journal of Operational Research. — 1992. — Т. 58, № 2. — С. 293—300.
4. Bertolini L., Le Clercq F., Kapoen L. Sustainable accessibility: a conceptual framework to integrate transport and land use plan-making. Two test-applications in the Netherlands and a reflection on the way forward // Transport policy. — 2005. — Т. 12, № 3. — С. 207—220.
5. Big Data defintion, Gartner Inc. — URL: `http://www.gartner.com/it-glossary/big-data/`.
6. Caprara A., Fischetti M., Toth P. A heuristic method for the set covering problem // Operations research. — 1999. — Т. 47, № 5. — С. 730—743.
7. Church R., Velle C. R. The maximal covering location problem // Papers in regional science. — 1974. — Т. 32, № 1. — С. 101—118.
8. Chvatal V. A greedy heuristic for the set-covering problem // Mathematics of operations research. — 1979. — Т. 4, № 3. — С. 233—235.
9. Costa Á., Markellos R. N. Evaluating public transport efficiency with neural network models // Transportation Research Part C: Emerging Technologies. — 1997. — Т. 5, № 5. — С. 301—312.
10. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters // Communications of the ACM. — 2008. — Т. 51, № 1. — С. 107—113.
11. Demchenko Y., Ngo C., Membrey P. Architecture framework and components for the big data ecosystem // Journal of System and Network Engineering. — 2013. — С. 1—31.

12. Demetsky M. J., Lin B. B.-M. Bus stop location and design // Journal of transportation engineering. — 1982. — T. 108, TE4.

13. Edmonds J. Covers and packings in a family of sets // Bulletin of the American Mathematical Society. — 1962. — T. 68, № 5. — C. 494—499.

14. Furth P., Rahbee A. Optimal bus stop spacing through dynamic programming and geographic modeling // Transportation Research Record: Journal of the Transportation Research Board. — 2000. — № 1731. — C. 15—22.

15. Gantz J., Reinsel D. Extracting value from chaos // IDC iview. — 2011. — T. 1142. — C. 1—12.

16. Garey M. R., Johnson D. S. A Guide to the Theory of NP-Completeness // WH Freemann, New York. — 1979.

17. Gleason J. M. A set covering approach to bus stop location // Omega. — 1975. — T. 3, № 5. — C. 605—608.

18. Haddadi S. Simple Lagrangian heuristic for the set covering problem // European Journal of Operational Research. — 1997. — T. 97, № 1. — C. 200—204.

19. Handy S. L., Clifton K. J. Evaluating neighborhood accessibility: Possibilities and practicalities // Journal of transportation and statistics. — 2001. — T. 4, 2/3. — C. 67—78.

20. Hine J., Mitchell F. Transport disadvantage and social exclusion: exclusionary mechanisms in transport in urban Scotland. — Ashgate Publishing, Ltd., 2003.

21. Jansson K. Optimal public transport price and service frequency // Journal of Transport Economics and Policy. — 1993. — C. 33—50.

22. Karlaftis M. G. A DEA approach for evaluating the efficiency and effectiveness of urban transit systems // European Journal of Operational Research. — 2004. — T. 152, № 2. — C. 354—364.

23. Kitchin R. The real-time city? Big data and smart urbanism // GeoJournal. — 2014. — T. 79, № 1. — C. 1—14.

24. Lam C. Hadoop in action. — Manning Publications Co., 2010.

25. Lan G., DePuy G. W., Whitehouse G. E. An effective and simple heuristic for the set covering problem // European journal of operational research. — 2007. — Т. 176, № 3. — С. 1387—1403.

26. Levinson H. S. Analyzing transit travel time performance. — 1983.

27. Murray A. T. Strategic analysis of public transport coverage // Socio-Economic Planning Sciences. — 2001. — Т. 35, № 3. — С. 175—188.

28. Murray A. T. A coverage model for improving public transit system accessibility and expanding access // Annals of Operations Research. — 2003. — Т. 123, 1-4. — С. 143—156.

29. Murray A. T., Davis R., Stimson R. J. Public transportation access // Transportation Research Part D: Transport and Environment. — 1998. — Т. 3, № 5. — С. 319—328.

30. O'Sullivan D., Morrison A., Shearer J. Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach // International Journal of Geographical Information Science. — 2000. — Т. 14, № 1. — С. 85—104.

31. Roth R. Computer solutions to minimum-cover problems // Operations Research. — 1969. — Т. 17, № 3. — С. 455—465.

32. Saka A. A. Model for Determining Optimum Bus-Stop Spacingin Urban Areas // Journal of Transportation Engineering. — 2001. — Т. 127, № 3. — С. 195—199.

33. Sampaio B. R., Neto O. L., Sampaio Y. Efficiency analysis of public transport systems: Lessons for institutional planning // Transportation research part A: policy and practice. — 2008. — Т. 42, № 3. — С. 445—454.

34. Shvachko K., Kuang H., Radia S. The hadoop distributed file system // Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. — IEEE. 2010. — С. 1—10.

35. Tao S., Corcoran J., Mateo-Babiano Exploring Bus Rapid Transit passenger travel behaviour using big data // Applied Geography. — 2014. — Т. 53. — С. 90—104.

36. The Big Data Long Tail. — URL: http://www.devx.com/blog/the-big-data-long-tail.html.

37. Toregas C., Swain R., ReVelle C. The location of emergency service facilities // Operations Research. — 1971. — Т. 19, № 6. — С. 1363—1373.

38. Transport Q. Integrated regional transport plan for South East Queensland // Government of Queensland, Brisbane. — 1997.

39. White P. R. Public transport: its planning, management and operation. — Routledge, 2008.

40. White T. Hadoop: The definitive guide. — "O'Reilly Media, Inc.", 2012.

41. Данные административного деления Торонто. — URL: http://opendata.toronto.ca/gcc/neighbourhoods_planning_areas_wgs84.zip.

42. Демографические данные районов города Торонто. — URL: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods.

43. Дрючин Д. А., Майоров М. А. Основные направления повышения качества транспортного обслуживания населения городским пассажирским транспортом по регулярным маршрутам // Вестник Оренбургского государственного университета. — 2015. — № 4. — С. 179.

44. Официальный сайт геоинформационной системы QGIS. — URL: http://www.qgis.org/en/site/.

45. Официальный сайт прикладного интерфейса Overpass API. — URL: http://overpass-api.de/.

46. Платформа данных Hortonworks. — URL: http://hortonworks.com/products/hdp/.

47. Портал GTFS Data Exchange. — URL: http://gtfs-data-exchange.com/agency/ttc/.

48. Портал OpenStreetMap. — URL: http://openstreetmap.org.

49. Прикладной интерфейс Hadoop Streaming API. — URL: https://hadoop.apache.org/docs/r1.2.1/streaming.html.

50. Программная обертка интерфейса Overpass API на языке программирования Python. — URL: https://github.com/mvexel/overpass-api-python-wrapper.

51. Спецификация GTFS. — URL: https://developers.google.com/transit/gtfs/.

52. Страница облачного сервиса Microsoft Azure HDInsight. — URL: https://azure.microsoft.com/en-us/services/hdinsight/.

53. Страница проекта Apache Hadoop. — URL: http://hadoop.apache.org.